Mine Automation and Data Analytics

Prof. Radhakanta Koner

Department of Mining Engineering

IIT (ISM) Dhanbad

Week - 7

Lecture 33: Introduction - I

Welcome back to my course, mine automation and data analytics. For the last few lessons, you have seen different automation technologies adopted in the mining industry and how using automation mining achieves higher productivity, safety, and efficiency. Also, you probably noticed that this system comprises several or multiple sensors in this process, and this sensor data is an extensive database for establishing control over the process. Also, this data is a big help for identifying the fault error in the process. In summary, the advent of this technology and installations and integration of different sensing systems in the mining process have given rise to a massive volume of data. So, how will this data benefit the mining industry needing data analysis? So, the primary data analysis process we can think of is the statistics. So, for the following few lessons, we will go through the different integrations of the statistical methods that need to be applied to the data, and using that technique, we are helping to analyze the data and understand the data, and that is our end objective to find out the trend from the data and by finding out the trend we want to optimize the process.

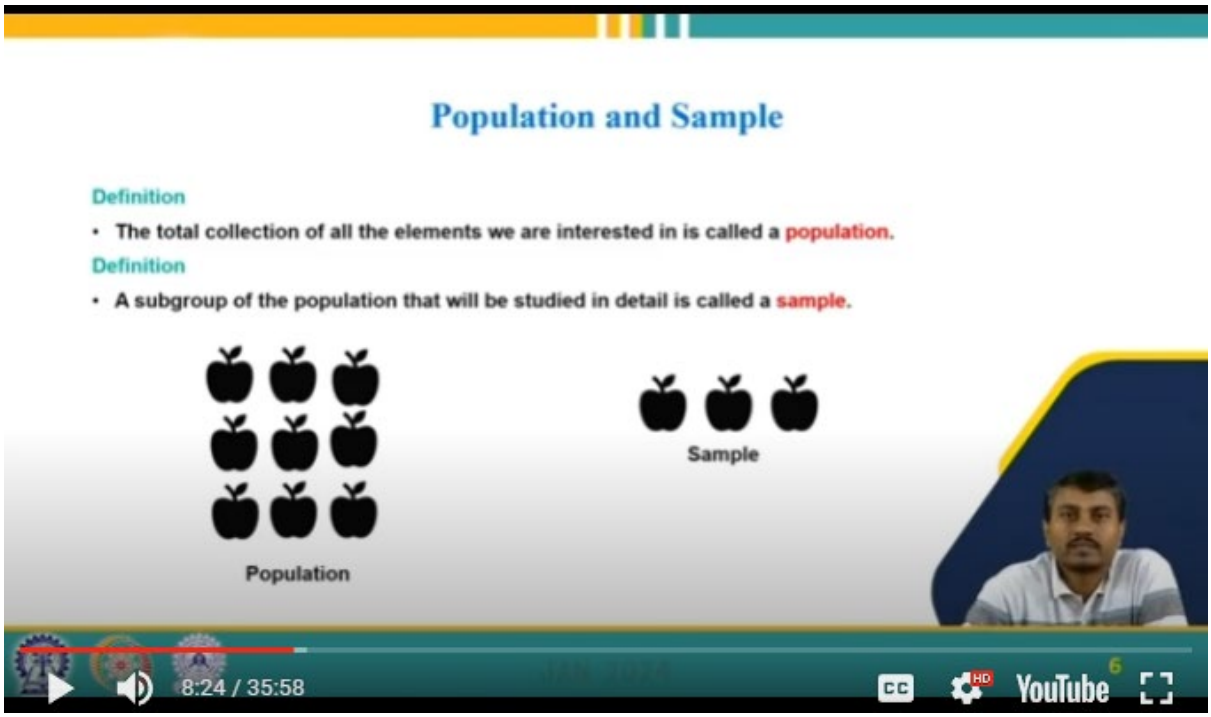
So, let's focus on the statistical aspect or technique that needs to be applied to this data. So, under module 7, this lecture would be the introductory part of the statistics. So, in this lesson, we will cover the following. The types of data, the classification of the kind of data between the numerical and the categorical, and then a visual representation of numerical data and interpretation of the shape of the distribution.

Then, we will discuss how to compute and interpret numerical data. Under this, we will discuss the central tendency, mean median mode, and the measures of dispersion, that is range. The other measures of dispersion we will discuss in the next lesson. So let's see what is statistics. So statistics is the art of learning from the data. Statistics says something about the process, the phenomena, and the process.

So, it is concerned with the collection of data, its subsequent description, and its analysis, which often leads to the drawing of a conclusion. So statistics is all about a process we have to follow for data collection and preparation, and then based on the data, we have to analyze, and by studying, we want to draw some conclusions. So, ultimately, the statistical analysis's result is to reach a fruitful conclusion. So, there are two major branches of statistics. One is the descriptive statistics, and the other is the inferential statistics.

So, what is descriptive statistics? Descriptive statistics is the part of the statistics concerned with the description and summarization of the data. Different kinds, for example, how many times the accident took place, okay? What is the total production, and what is the increase in production? So, based on these data descriptions, there are different descriptions of the health of the method. Another is the inference. The part of the statistic concerned with concluding data is called inferential statistics.

So now, based on the data, we want to draw a new conclusion. In that process, when we follow, we have to take care of the inferential statistics. So when we are going to draw some decision or conclusion from the data, we have to consider the possibility of chances because these are the random data we are collecting, and based on the data, there is a chance that this is the possibility that this will come this number of times. So, that gives rise to the new concept of probability; we will discuss this in a subsequent lecture.



Population and sample. So, in statistics, these two things are essential, and we have to understand what population is and what a sample is. So the population is the whole set total population, a total set of data we have collected about a process, okay, and we are interested in the process to know what is happening so when all the data comes together, that comprises the population and when you want to draw some amount of conclusion out of the data because when the data size is significant, we might not be able to accommodate in our analysis process all the data. So we will take some of the data out of it so some of the data we are collecting from the population is a half group that is a sample. So, this sample is collected for study and analysis to conclude this population.

So, the sample is a subgroup of the population. So this is the larger population group's left side, and a part of it is the sample. So, the sample is always smaller than the population—the purpose of statistical analysis. We might have to use this statistical technique for several purposes.

One is if the statistical purpose of the analysis is to examine and explore information for its intrinsic interest only the study is descriptive, and if the information is obtained from a sample of a population and the purpose of the study is to use this information to conclude the population the analysis is inferential. So, descriptive research may be performed on a sample or population when an inference is made about the population based on the sample information, and then the study becomes inferential. So let us see what data is because here, data is all about. The statistical method concentrates on the data; you have to generate and collect data. So, what is data? Data represents something about the natural process of phenomena, such as processes like that.

So, to learn something, we need information. So, the data are the facts and figures collected, analyzed, and summarized for presentation and interpretation. So, in computer graphics, you can see that infinite data is generated due to the simulation of different computation processes. So, there is another kind of data graph plot, histogram, which represents some status of a process. This is also a data.

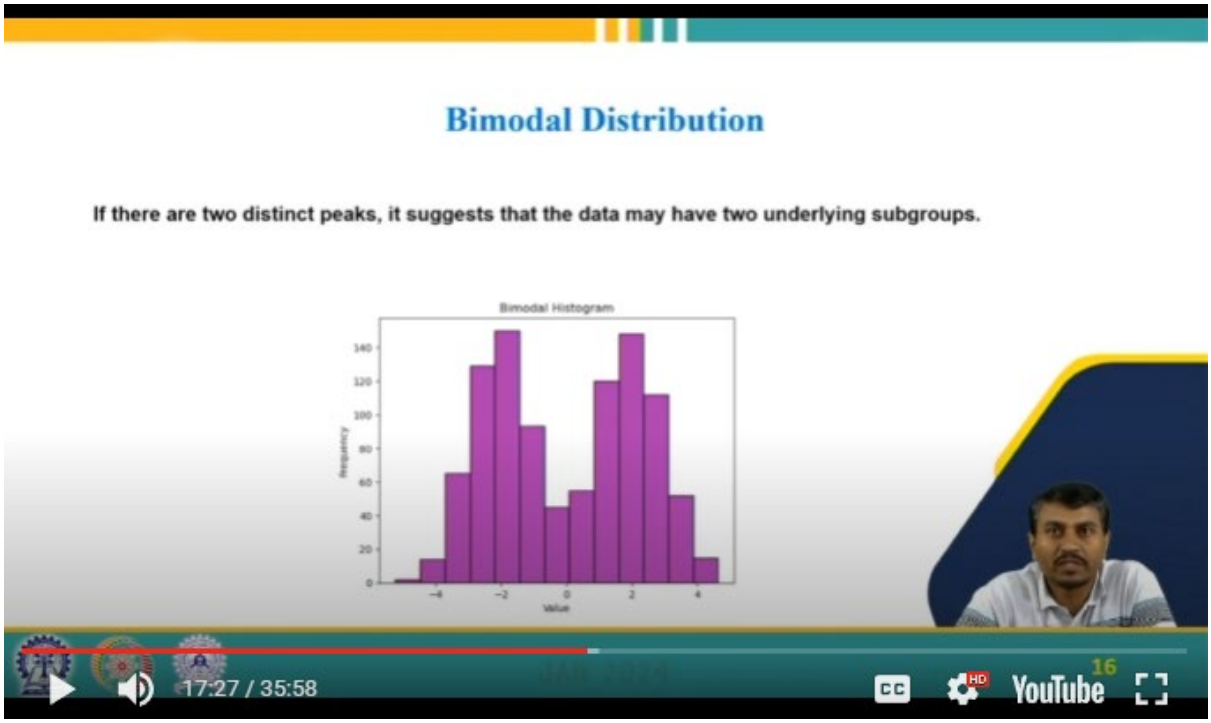
Variables and cases: Another critical concept we must understand is variables and cases because of the number of times we use the term variables. So, what is a variable? So, a variable varies over the process, okay over the process chain, and over the process system, so it varies. In formal definition, it is a characteristic or attribute that varies across all units, and the case is the observation of a unit from which data is collected. So, the row represents the case for which the same attribute is recorded, or each case, and the column represents the variable where the same type of value for each case is recorded for each variable. So here is the sample data: serial one and serial two, then the student name Radha, Krishna Sai Raj, English mark is 93, 89, science mark is 85, 91. So here, in this column, the student name is the variable, and serial number wise for Radha Krishna this is serial number one is case one, two is a Sai Raj, other case two, so this is the case and observation case or observation.

So, each variable must have its column, and each observation must have its row—classification of data categorical and numerical. So, there are different types of data available in different scenarios. For example, in a mining case, if you analyze, several data might come in the process, for example, the different production-related data, such as the number of tons or fuel consumption, that is also the number of liters okay, and also the energy efficiency in percentage that is also numerical data. The quality of the categorical data in the environment may be excellent and moderate. This kind of situation is moderately clumsy weather or dusty situation very dust so these are categorical data, and we want to apply statistics first on the numerical data; there are different processes that on the categorical data you can also

nomenclate in the numerical terms you can you can subdivide the category into different groups and different range and based on that you can now put an artificial scale it into a numerical data as well for your analysis based on the situations. The first thing in the statistical process is the visual representation. You need to represent the data meaningfully, which might be helpful for the user.

The visual representation of numerical data is a crucial aspect of data analysis and provides insight into the distribution and characteristics of the data. The histogram is one of the standard methods for visualizing numerical data, and this is how different frequencies of a value have been plotted here. This is another kind of histogram. The middle one shows a histogram of data. This is the histogram showing a negative trend, this direction, a positive trend, and this direction. So, a histogram is a graphical representation of the distribution of a data set. Hence, it consists of bars where each bar represents a range of values called a beam, and the height corresponds to the frequency of observation within that beam. So, for example, here is the data set 1 2 3 4 4 4 5 5 6 6 6 6 7 7 8 8 8 9 9 9 9 so this is now categorized into 1 2 3 4 5. Number of beams and their frequency is plotted. So, this is a typical histogram plot of the data.

Symmetric distribution: if the histogram is roughly symmetrical, it suggests the data is evenly distributed around the mean. So, this is one symmetrical histogram data representing that the data is symmetrically distributed over a mean. Skewed distribution, if the histogram is skewed to the right, indicates the data has a tail on the right side. Okay, another is similarly skewed to the left, which is negatively skewed. The tail on the left side is a left-skewed histogram. This is the right-skewed histogram, representing some skewness in the data. This is not evenly distributed, and meaning is not in the middle. If there are two distinct peaks, bimodal distribution suggests the data have two underlying subgroups, so these are the two sharp peaks that represent two subgroups within the data set.



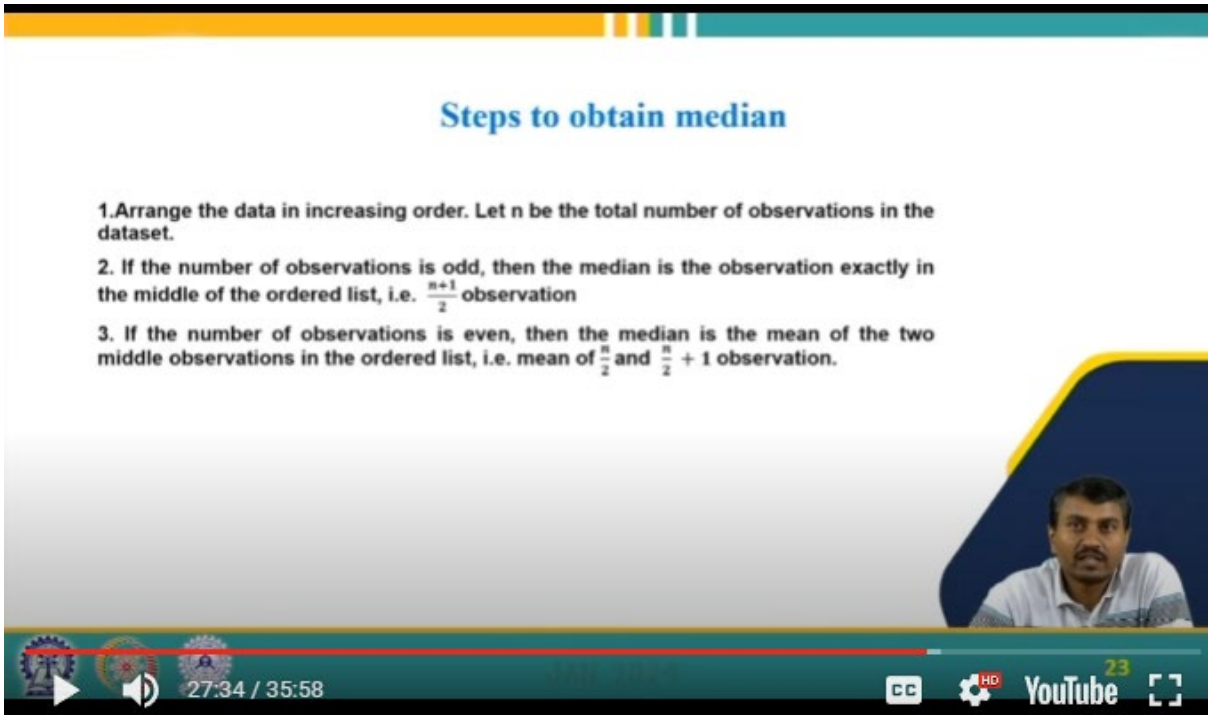
Descriptive measures are the most commonly used and can be categorized as measures of a central tendency. These measures indicate the most typical value or center of a data set, and the measures of dispersion indicate the variability of a data set around the mean. For measures of central tendency, there are three terms: mean median mode, and for example, this is the data set 1 2 3 4 4 4 5 5 6 6 6 6 7 7 8 8 8 9 9 9 9 9 so here it is the histogram that is plotted for this data set now the mean is average of all these data summing over all these data numbers of observation divided by that so X1 X2 X3 up to Xn divided by n that is the mean the median is the 50% value this side and 50% value this side the middle value so here you can see if you arrange this is already placed in ascending order so here several observation is 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 so it is 11 and 12 in middle of 11 and 12 1 2 3 4 5 6 7 8 9 10 11 12 so 6 and 6 average is six medians are six and the mode the highest number of occurrence 1 2 3 4 5 5 number of times nine has occurred so the mode is the 9

the mean the most commonly used measures of central tendency is the mean the definition the mean of a data set is a sum of the observation divided by the number of observations so the mean is usually referred to as average for discrete observations sample mean is X bar X bar is the representation of the mean of the sample is equal to X1 X2 ... Xn divided by n sample population mean is the mu so X 1 X 2 X n divided by the capital N population sizes bigger that is capital N so this is a typical data set and its mean is demarcated using these red dotted lines the example so here we have tried to show you an example that the winning score of the US master golf tournament in the year 2004 to 2030 were 280 278 272 276 281 279 276 281 289 and 280 so find the sample mean of these course so these are basically added and number of observation is 10 so the mean is basically 279 point 2 adding multiplying a constant so the thing is so when you add some value on the sample okay a constant value for all these values then the mean is also added by that amount so the mean of these new sample space is basically the old mu plus the C how it is happening so for example X 1 was the first data then X 2 is the second data dot dot dot dot dot we have X n number of data so the mean was X bar equal to X 1 plus X 2 dot dot dot dot X n divided by n now the new set is basically X 1 plus C X 2 plus C dot dot X n plus C the new mean of this sample space X bar new is equal to X 1 plus C plus X 2 plus C dot dot X n plus C divided by n number of parameter is n so X 1 plus X 2 dot dot dot dot X n plus n into C because n number of times he occurs divided by n so this part is nothing but the old X bar X bar plus n C divided by n is equals to X bar plus C so the new mean of these set is basically the old mean plus the C

So for the multiplication for the same data if we apply the multiplication here the mean will be the the old mean multiplied by the multiplication factor how it is happening similarly so X 1 X 2 dot dot dot dot X n is basically the data set now the X bar is equal to X 1 plus X 2 plus X n divided by n so the new set is equal to C X 1 C X 2 dot dot dot dot C X n so the new mean of this data set is basically X new is equal to C X 1 plus C X 2 plus dot dot dot dot C X n divided by n so is equal to C is a common factor so now X 1 plus X 2 dot dot dot X n divided by n so this is nothing but this particular part divided by n is nothing but C into X bar so when we multiply some constant factor to the sample data the mean is also shifted to multiply that by that factor and also when we add constant factor on the data it basically shift with add to

the old mean with that amount of constant so here is the example marks of the student has been given 88 79 74 86 87 67 90 and 49 so the mean of the marks is 75.

67 now let us suppose you have decided to add 5 marks as a bonus to each student so the data will become like this and the mean of the new data set is nothing but 80.67 that is 75.67 plus 5 the median another frequently used measures of center is the median and especially the median of a data set is the number that divides the bottom 50 percent of the data from the top 50 percent so the medium of a data set is the middle value in its ordered list so this is basically a histogram plot from that you can see the median so how to obtain a median from the data so first we have to arrange the data in increasing order and let n be the total number of observation in the data set now we have to remember if the number of observation is odd then the median is the observation precisely the middle of the order list that is n plus 1 divided by 2 observation if the number of observation total is basically even then the median is the mean of the two middle observation in the order list that is n by 2 and n by 2 plus 1 observation so if the sample space value is 22 so it is basically 11 and 11 plus 1 12th observation so average of the 11th and 12th observation is the medium of that sample



example of median so here is the score we have 75 82 90 65 88 72 91 78 85 and 79 and the median is 80.5 When we ordered these in increasing order, none of the values were found, so this is 1 2 3 4 5 6 7 8 9 10; it is an even amount of data. Ten means five plan and six, so the average of the 5th and six are coming as 80.5, another number 11 22 15 29 33 15 17 22 19 25 and 27 the median is 22 1 2 3 4 5 6 7 8 9 10 11 so exactly 11 means 11 plus 1 12 divided by 2 6 6th observation is the 12. Hence, the median has a 12 outlier effect, a fundamental concept to understand. For example, the example is 12 12 12 5 7 6 7 3, and the mean is 6. It is 1 2. It should not be 12 1 comma two, so the sample mean is six. The sample median is also six, so now you add some error on the data in some out layers. For example, now it is 2 1 17 5 7 6 7

3. Now, in this data set, the mean is changed to 21. The sample mean is shifted, so when you add some outliers, the sample mean is affected, but the sample median remains unaffected.

So that is an essential observation so we can use this median filter so median filter is a outlier resistant so that is one good observation so the sample mean is sensitive to outlier whereas the sample median is not sensitive to outliers adding multiplying a constant so here for the same if the same rule applies as we have seen for the mean that the new median when you multi when you add some value constant value over the sample space the new median will be the old median plus that constant value similarly when you multiply some continuous factor in the sample space the new median value is the old median value multiplied by the C mode another measures of central tendency is the sample mode the definition the mode of a data set is its most frequently occurring value highest number of times the value occurring in a sample space is the mode now how to obtain that value so what are the steps see if no value occurs more than once then the data set has no mode that is clear else the value that occurs with most significant frequency or the highest level of frequency or highest number of times is the mode of the data set for example 2 12 5 7 6 7 3 here 7 occurs twice so the mode is 7 but for this data 2 105 5 7 6 3 all the data occurs only once so there is no mode so this is basically good education that some of the data set might not have mode adding multiplying a constant same effect when you add some continuous value in the sample space the mode is basically the new mode is basically is the old value plus

The constant value similarly when you multiply C it is the old value multiplied by the C factor now the relationship of mean median and mode so in statistics for a moderately skewed distribution remember so it is an empirical relation though so there exists a relation between mean median and mode and this mean median mode relationship is known as empirical relationship which is defined as mode is equal to the difference between three times the median and the two times the mean so mode is equal to 3 median minus 2 mean measures of dispersion to measure the amount of variation are the spread in a data set it basically comes under the study of measures of dispersion so measures of dispersion are also known as measures of variation or spread across the range so measures of dispersion range variance standard division and inter quartile range so what is range so the range of a data set is the difference between its most significant and most minor value so range of a data set is the difference between the highest value and the smallest value so this range is essential to know the data set spread so range is basically max minus mean so where the max and mean denotes the maximum and minimum observation respectively in the whole data set so range is sensitive to outlier because

The max value might be the outlier value, so it is sensitive. Even the mean value might be the outlier value, so it is susceptible to the outlier. These are the references, so let me summarize what we have covered in this lesson in a few sentences. We have learned the types of data and identified the data as categorical and numerical. We have learned the visual representation of the numerical data and interpreted that the distribution's shape is symmetric, skewed, and bimodal. We have computed and analyzed the numerical summaries of the data. Under that, we

have calculated the mean median and mode, and under the measures of dispersion, we have learned how to calculate the range. Thank you.