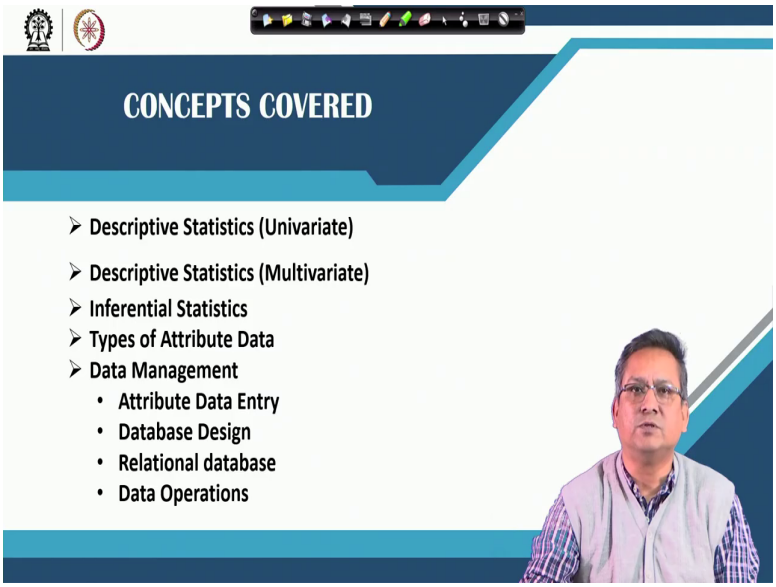
Geo Spatial Analysis in Urban Planning
Prof. Saikat Kumar Paul
Department of Architecture and Regional Planning
Indian Institute of Technology, Kharagpur

Lecture – 11
Attribute Data Management and Data Exploration

Welcome back dear students. We are into the module 3 now and we shall talk about Attribute Data Management and Data Exploration in this particular lecture.

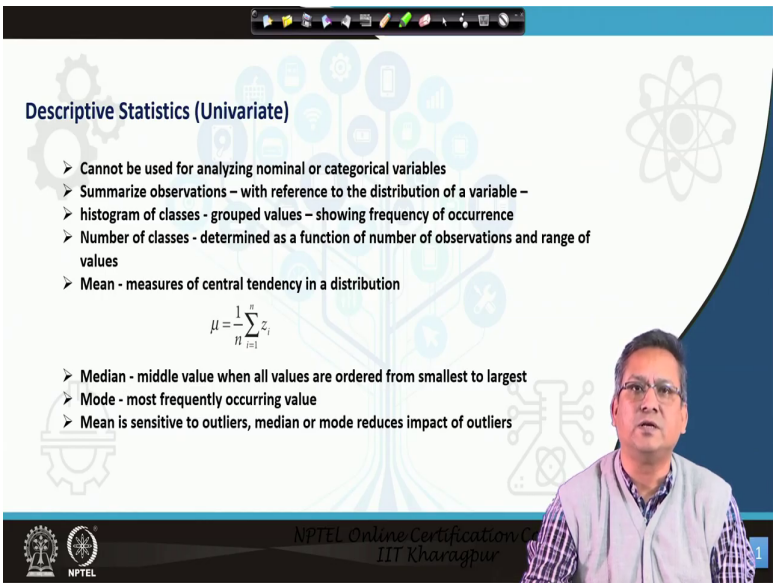
(Refer Slide Time: 00:27)



So, the concepts that we are going to cover today is about Descriptive Statistics and we are going to talk about the Univariate Statistics, we are also going to talk about touch upon Multivariate Statistics, we would see what is Inferential Statistics, we shall also look into the types of Data Attributes and how the data is managed in GIS.

So, how we do the Attribute data entry, how we do the Database design, what is a Relational database and the I mean different types of data operations regarding joining of the tables. So, we shall look into these following concepts.

(Refer Slide Time: 01:09)



So, talking about Descriptive or Univariate statistics, it is used I mean for the ordinal variables and the ratio variables and it cannot be used for nominal or categorical variables. We shall see what are these nominal and categorical variables. Some of you might be knowing about these who are into I mean data management and I mean database design and all these things.

So, it basically summarises the observation with descriptive statistics, it gives a summary of the observations and it is with reference to the distribution of a variable. So, whenever we have variables and we have distributed data or discrete data, it is lumped or grouped into

different classes or grouped values and we have frequencies of occurrence for each of these groups which can be shown as histogram of the different classes.

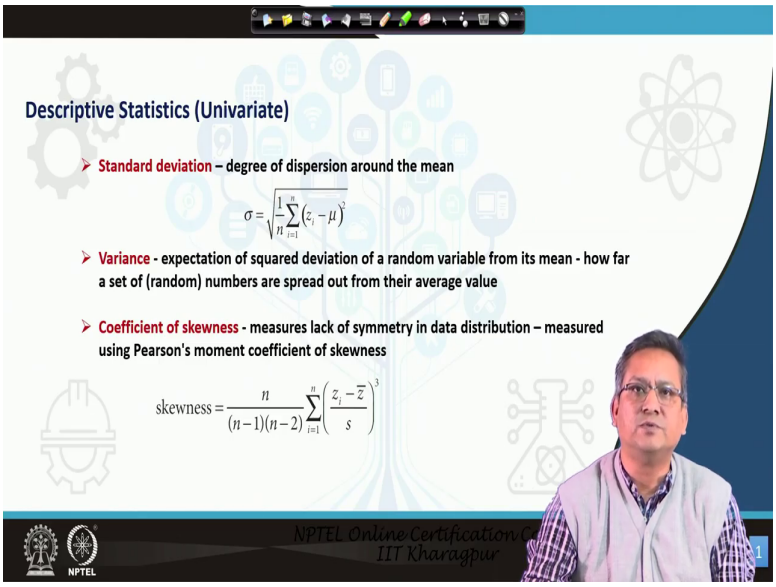
Now, I mean the number of classes that we would categorise our data is determined as a function of the number of observation and the range of the values. Now, talking about the measures of central tendencies, we have three measures which you already know about. The first one is known as Mean which is I mean the average of all the sets of observation.

So, if we have n number of observations, we sum up the I mean observations all the "n" observations and divide it by the number of observation to give us the mean which is denoted as mu. Now the next measure of central tendency is the Median which is the median middle value when all values are ordered if we I mean stack up the values from the smallest to the largest.

So, you can find out the middle value and in case if you have even number of data sets, in that case what we do is we take the two middle values and I mean take mean of those. Now, you have different types of algorithms in computer I mean software analysis. So, we can I mean categories the data or I mean save the data in a I mean smallest to largest I mean ordering. So, those different algorithms are available where in if we have a very large data set, it can be ordered and then we can find out the median.

Now, mode is the most frequently occurring value in a data set. So, whatever data value is the most frequently occurring data value I mean that we take it as a mode. Now, if we have outliers in the data, if we have some data values which are extreme data values I mean which is beyond the I mean central any of the central tendencies and which has very high values, so in those cases what happens this outliers will add up significantly to the mean. So, in a way the mean gets I mean affected by this kind of outliers. So, mean median or mode value it reduces the impact of the outliers.

(Refer Slide Time: 04:52)

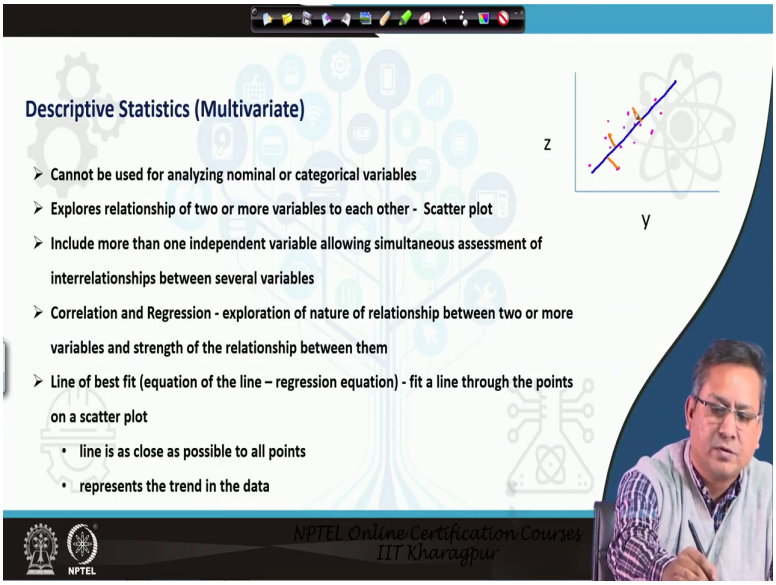


So, I mean we talk about the descriptive statistics. So, we have three measures basically we talk about the standard deviation where in we have the degree of the dispersion around the mean and it is represented as sigma which is the difference of the square I mean of your observed values to the mean are divided by the total number of observations and it is taken as a square root.

We also have a term which is known as Variance which is I mean extensively used in GIS or otherwise in descriptive statistics. So, I mean it is the expectation of square deviation when the variable is random with respect to its mean, now I mean it gives us a spread out of those random numbers with respect to the average value with respect to the mean values. Talking about the coefficient of skewness when we are having a distribution a, when we are having a histogram we can see that it could be in the form of a inverse bell curve.

So, this curve could be symmetrical in nature or it could be asymmetrical in nature. So, the coefficient of skewness, it measures the lack of symmetry or the presence of symmetry in the data distribution and it is measured using Pearson's moment coefficient of skewness which is given by this particular equation shown here.

(Refer Slide Time: 06:32)



Now, going to Multivariate Statistics, generally I mean we are dealing with data which is not univariate, but it is multivariate where in we have one dependent variable and we may have multiple independent variable.

So, I mean your univariate statistics cannot be used to analyse such situation. So, I mean again when we like we were talking about univariate data in case of multivariate analysis as well I mean it is not possible for us to analyse the nominal or categorical variables, so we can as I

was telling that we can explore if there is a relationship between two or even more than two variables.

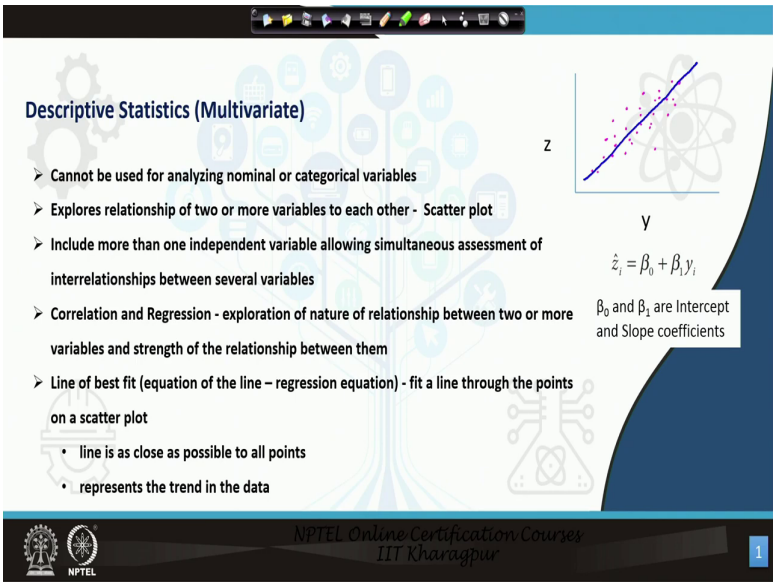
So if you have two variables, then it would come as a scatter plot you can have abscissa and ordinate x-axis and y-axis and the data plots would data values would be scattered in as paper pot, I mean diagram as I mean as dots in the x and the y axis. So, I mean if you have a three dimensional data set in which you may have one independent, dependent variable and two independent variable, in that case it would look like a cloud of points and similarly, if you have more number of variables it becomes a figure becomes complicated.

So, I mean we can include more than one variable and but the advantage of this multivariate data descriptive statistics is that we can simultaneously assess the interrelationship between multiple number of variables. Now, to do this assessment what we do is, there are two methods. We are talk about the correlation and we talk about regression.

So, in correlation and regression we nature I mean we explore the nature of relationships between the different variables and we try to assess a what is the strength of relationships between these variables. So, when we have these variables, what we can do is we can fit a line to these data points and this line is the line of the best fit, that is the line which passes through this scatter plot is worked out. So, it is as close as possible to all the points and it represents I mean this particular line or the best fit line it represents the trend in the data.

So, suppose we have this data points and we I mean try to create this data points scatter plots in this particular x, y and z values two variables which is y and z. So I mean if we try to do a best fit plot, then what happens is you can find out the distances and these distances would be minimised, these distances from this points would be minimised, so that the this is the best fit line.

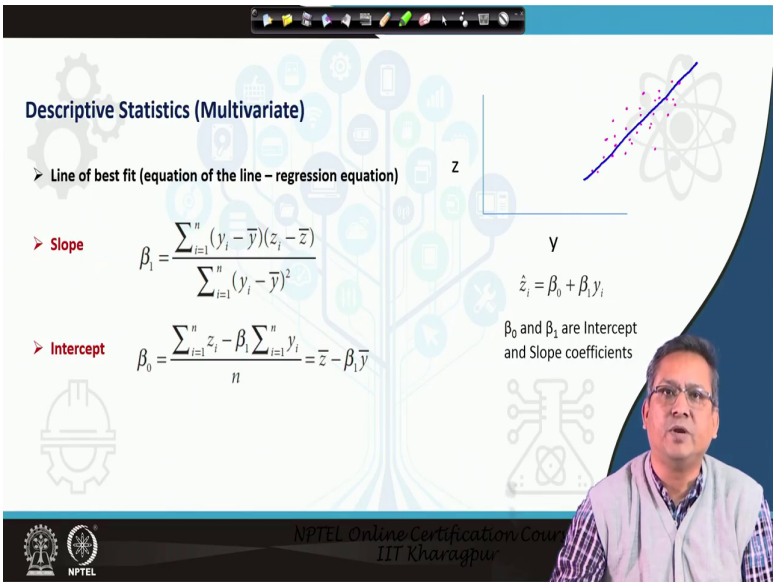
(Refer Slide Time: 10:47)



Now, if we see the equation of the best fit line, it is in the form of y equals mx plus c where in we have the slope and we have the I mean intersect. So, this in this particular equation that is z i equals to I mean beta subscript 0 plus beta subscript 1 y i in this beta 0 represents the slope and the beta 1 represents the intercept coefficients. So, we see this particular equation in this your z i is equal to beta 0 beta subscript 0 plus beta subscript 1 yi.

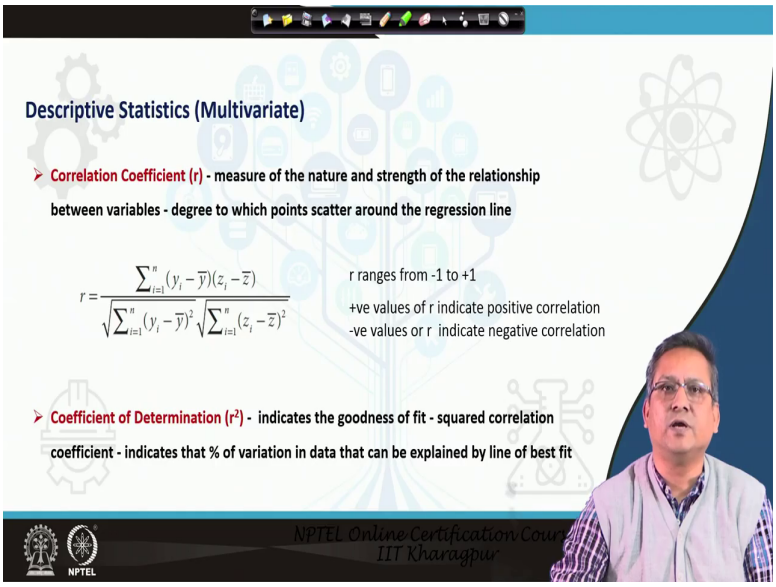
This is the equation of this particular line which is the line of best fit in which it is as close as possible to the to all the points. So, we shall see how we minimise this distance from all these points. So, in this case this is of the form of y equals to mx plus c. So, we see the beta subscript 0 and beta subscript 1 are the intercept and the slope coefficients.

(Refer Slide Time: 12:06)



Now, we had talked about the I mean the best fit line. So, this is the regression equation. This equation that we had talked about is the regression equation. Now talking about the slope of this particular line that is the beta 1 it is I mean calculated using this particular equation I mean it gives you the ratio of the rise along the x to the ratio of rise along the y axis. So, in this case it is given by this particular equation. So, similarly we can work out the calculate the intercept by this equation. Again it is I mean subtracting your z mean minus beta 1 into y mean.

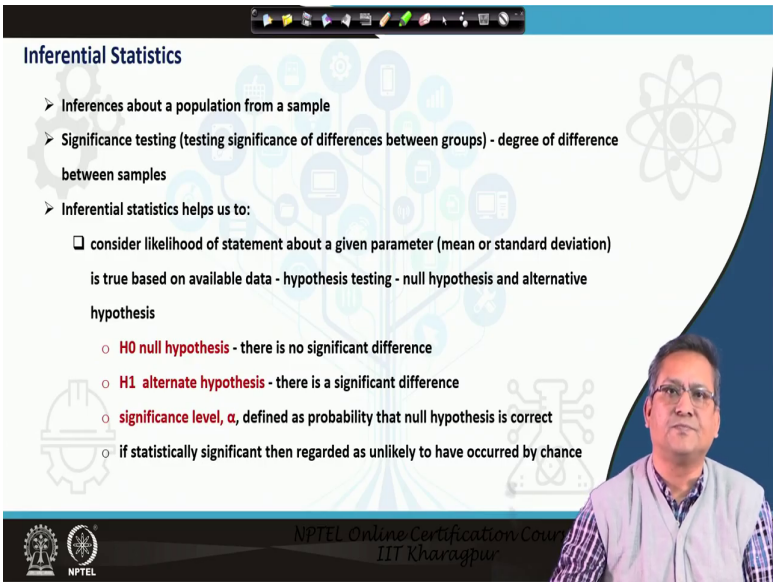
(Refer Slide Time: 12:55)



Talking about the correlation coefficient which gives us the strength and the nature of the relationship between the variables, so it gives us the degree to which the scatter points are around the regression line. So, it is a ratio again and it is worked out using this particular equation and we can see that the r that is the correlation coefficient, it ranges from minus 1 to positive 1 value.

Now, when we have positive values of r, it indicates a positive correlation and when we have negative r values as negative, it indicates we have negative correlation. Now, the coefficient of determination that is also known as r square or in some cases you will see there are values of adjusted r square, it would give the indication of the goodness of the fit. Now, it is the square of the correlation coefficient that we had seen earlier and it I mean gives us the percentage variation in the data that is explained by the line of the best fit.

(Refer Slide Time: 14:15)



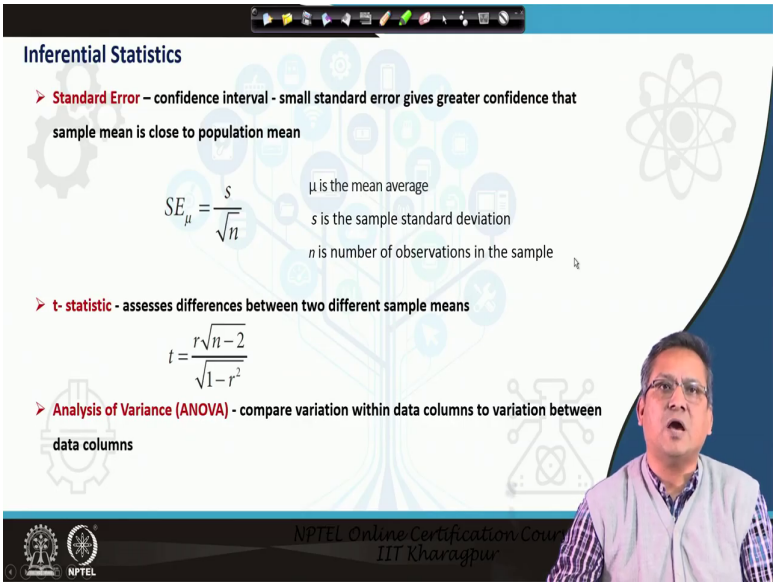
Now, talking about inferential statistics when we have a population huge population, we cannot do a sampling of the entire population. So what we can do is, we can take a sample of the population and that sample has to be significant. So what we do in inferential statistics, we tried to see or do the testing of significance. So, this testing of significance is between the different groups that is the we try to measure the degree of difference between the different samples in the population. Now, this inferential statistics it helps us to consider the likelihood of statement about a given parameter.

So, when we are talking about a hypothesis before doing a research, so we bank upon this inferential statistics to see whether the hypothesis is correct or it is incorrect. So, we talk about two cases in terms of hypothesis 1 is the null hypothesis and one is the Alternate hypothesis. So when we do this inferential statistics, we can conclude that the if there is significant difference between the two groups I mean between the statistical parameters of the

two groups that is the mean or the standard deviation if the two groups were significantly different for the population.

So it gives us the null hypothesis and the other one is your alternate hypothesis that is when the significant there is a significant difference in the population, it gives the alternate hypothesis and we also have the significance level which is denoted as alpha and it is defined as the probability that null hypothesis is correct and if it is statistically significant I mean it is unlikely that the observation or the sample would have occurred by chance.

(Refer Slide Time: 16:36)

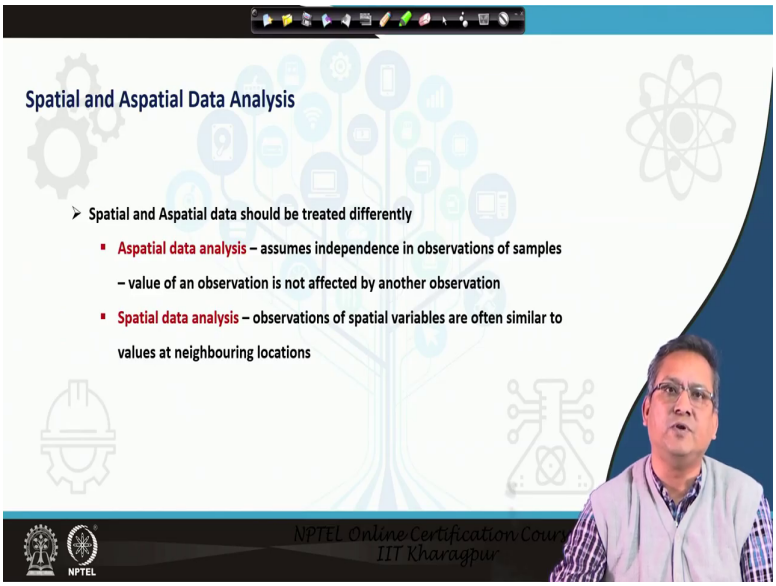


So, this is how we use the I mean or we can use your univariate statistics, multivariate statistics or inferential statistics. So, we have several other matrix in inferential statistics. First is the standard error where in we try to measure the confidence interval which is the I mean which gives us the sample mean and the population mean difference. So, if this error is small

that means the there is greater confidence that the two means are closer to each other. Now, the standard error is the ratio of standard deviation to the square root of the number of the observations in the sample.

We also have a measure which is known as t-statistic which assesses the difference between the two sample means. It is calculated using this particular equation which is r into root over n minus 2 divided by 1 minus r square. Now, we also analyse we can also analyse the variation within the data columns and between the data columns. So, it can be done using this particular method which is known as ANOVA which is analysis of variance.

(Refer Slide Time: 18:07)

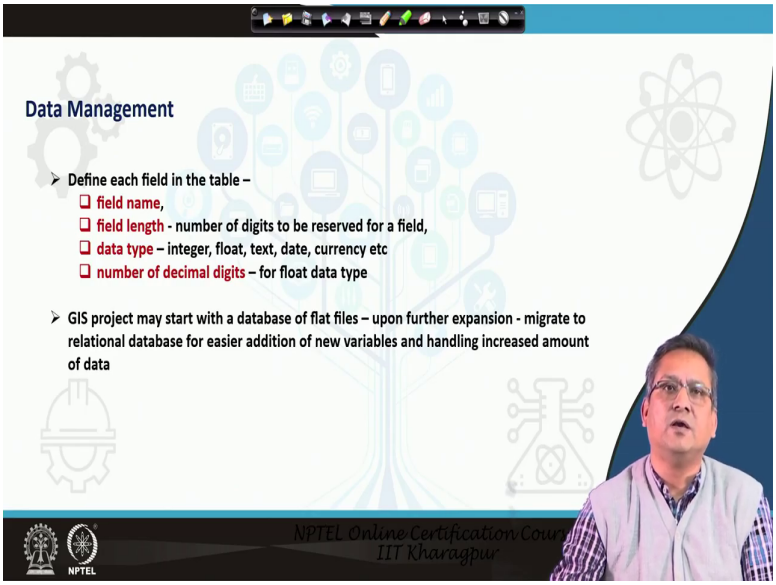


So, apart from this we can also I mean find out how your spatial data and aspatial data are different from each other. So, if we run the statistical inferences on spatial data, we can see that there is a difference between the spatial data sets and the aspatial data sets does. It should

be treated differently, in a way that your the premise when we are dealing with aspatial data, any sort of statistical analysis that we had talked about whether it is univariate, multivariate statistics in that case the assumption is that there is independence in the observation of the samples that is the value of an observation is not affected by the other observations in the given data set.

But when we are talking about spatial data analysis in real world, we would see that observations in the spatial variables in the spatial domain, they are often similar to the neighbouring values. So in a way when we are to analyse the spatial data, we should have a different approach, a distinctly different approach than the approach that we use for a spatial data analysis.

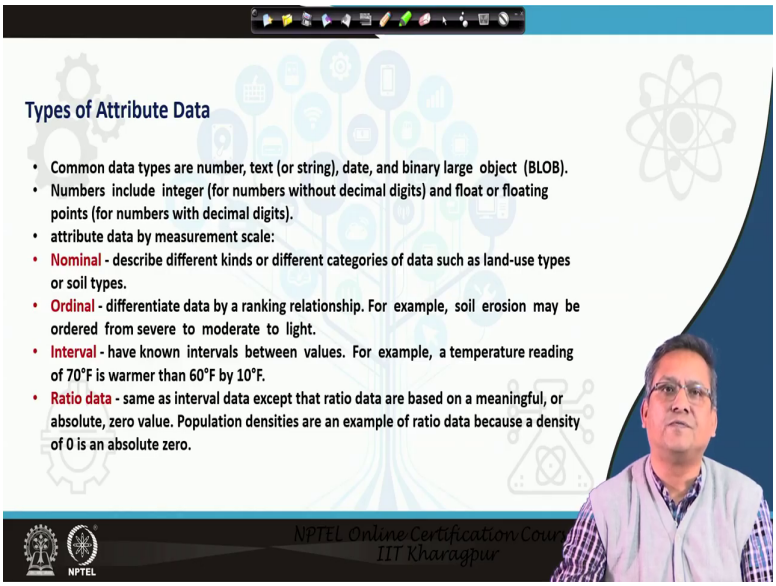
(Refer Slide Time: 19:39)



Now, talking about Data Management we can define the each field in the data table where in we give the field name, we can specify the field length which is the number of digits to be reserved for a field. We can also give the data type whether that data is integer, it is float or it is a text information, whether it is a date or a currency information and we can also give the number of decimal digits in case the data is a float data type.

So, I mean we when we are doing this analysis, initially we can start with flat files. So I mean when we have more amount of data, then we can get into a relational database. If we have a huge data set, we can create a relational database and we can it becomes easier to handle in such a case when we are using the relational database management for handling huge amount of data.

(Refer Slide Time: 20:51)



So talking about the different types of attribute data, we have the number text date binary or large binary large object which is also known as BLOB data. when we are talking about numbers, we would be dealing with either I mean real numbers, integer numbers. So, we can have integer values or we can have float values. So, when we are coding the data specially for the raster type of data, are we have to be very careful about what kind of data set we are creating because say suppose if you have a float operation and your data type you are defining it as integer, it would discard all the float values.

So, we have to be very careful about the I mean the kind of data that we are handling. So, specifically in case of raster. So, similarly we can have float data. So, the data resolution also would come into picture when we are dealing with raster data set. It could be a 2 bit data, it could be a 4 bit or 8 bit data, 12 bit data depending on the I mean the size of the I mean difference of the highest values and the lowest values.

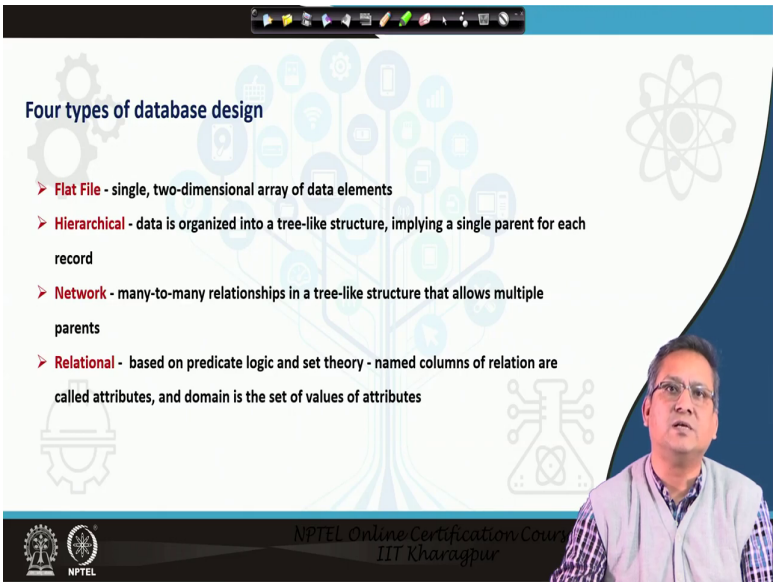
Now, attribute data is measured by different scales. So, we have different types of attribute data. The first one that we come across is the nominal data which describe the different kinds or categories of data. So, examples of this kind of data could be your land use data or data are such a soil data, so where in you give the categories of land use or the categories of soil. The next data that we deal with is the ordinal data. So, it I mean differentiates the data by a ranking relationship.

So, we can say suppose we have a data set, so we can quantify the intensity like in this particular case we are talking about soil erosion. So, whether the erosion is severe or moderate, we can also talk about proximity matrix like it is very near or far or I mean we can have different types of measures and rank the relationships. Now, the interval data are the intervals between different values have known intervals, like suppose we can categorise the temperature data into different groups and we can have group where in we can have a cold data temperature which are normal or comfortable and temperatures which are warmer.

So, we can categorise the data and these are known as interval data. The last one which we often used for data modelling in GIS is the ratio data and this is the most powerful tool which

we have. So, we can use both integer type of your numerical data or float type data when we work with ratio data sets. So, your I mean it is the ratio basically. So, I mean for an example we can talk about the population density in different words which is an example of ratio data.

(Refer Slide Time: 24:33)



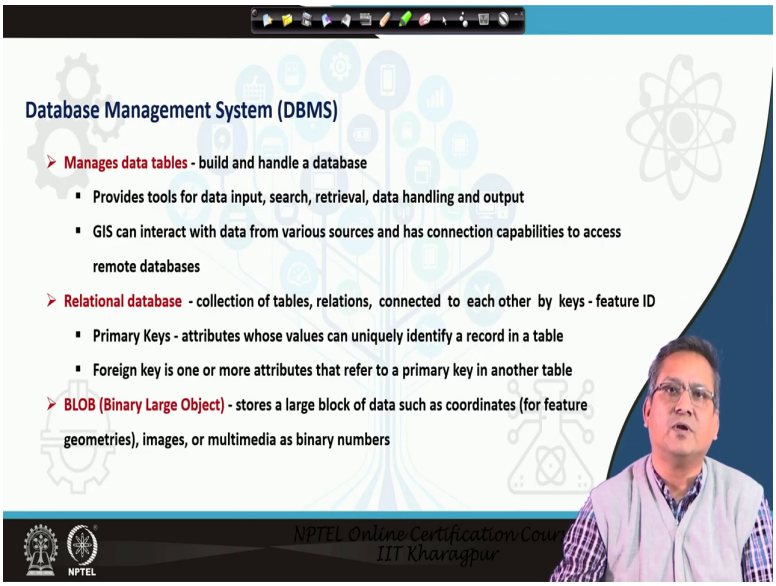
Now, there are four types of database design whenever we take GIS tasks, we can create different types of database depending on the size of the database. So, if the database is small we can create a flat file nomenclature where in we have a single file and the data is arranged in a two dimensional array of data elements. The next one is the hierarchical data in which the data can be organised in a tree like structure and it implies that there is a single parent for each record.

The next data type is the network data type. So, it is a modification, further modification of the hierarchical data set and it embodies many to many relationships in a tree like structure in a

hierarchical structure. So, this data structure or this type of database design it allows for multiple parents. So, if you have leaf nodes you can have multiple parents to those leaf nodes.

The last type for the database design is the relational database which is the most powerful data I mean database design when you have very big data sets. So, we generally used or relational database management system to model the entire data to do the designing of the entire data sets. So, I mean it is based on predicate logic and it is based on set theory. So, we have name columns of relation which are called attributes and domain and this domain is the sets of values of the attributes.

(Refer Slide Time: 26:27)



So, we can have database management systems which work in the back end when it handles your GIS data sets the attribute data. So, it builds and handles the GIS database. So, these DBMS tools they provide I mean solutions for data input for search and retrieval for data

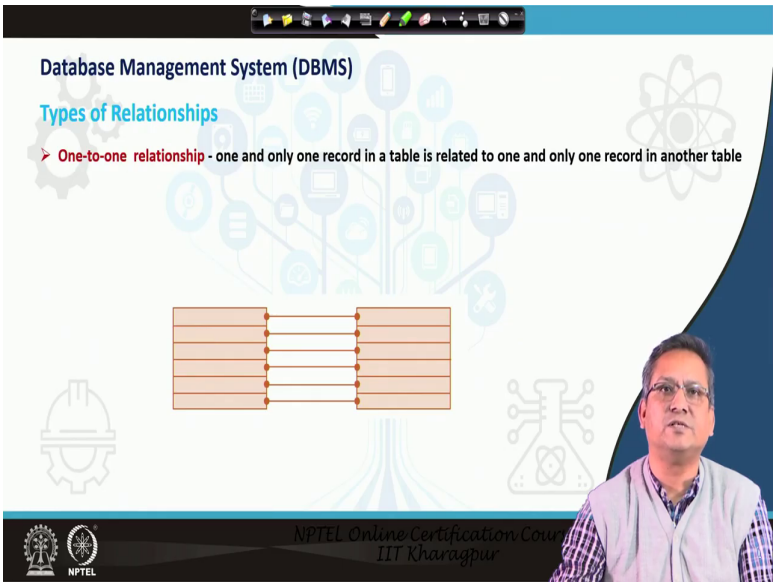
handling and for generating output from your queries, your GIS can interact from I mean data from multifarious sources.

So, I mean we can connect to remote data bases and this GIS has the capability to access such data bases I mean multiple data bases. So to I mean connect this multiple databases, we would need to have a unique field which is known as keys. So, we have a relational database where in we have collection of tables and they would be connected to each other by a feature id which is known as keys and the relations are built into it.

There are different types of keys. First is we talk about the primary keys I mean these are the attributes whose values are unique and it can be identified as a record in a table. We also have a foreign key which is one or more attribute that refer to the primary key in another table, another reference table in which some other data sets are available.

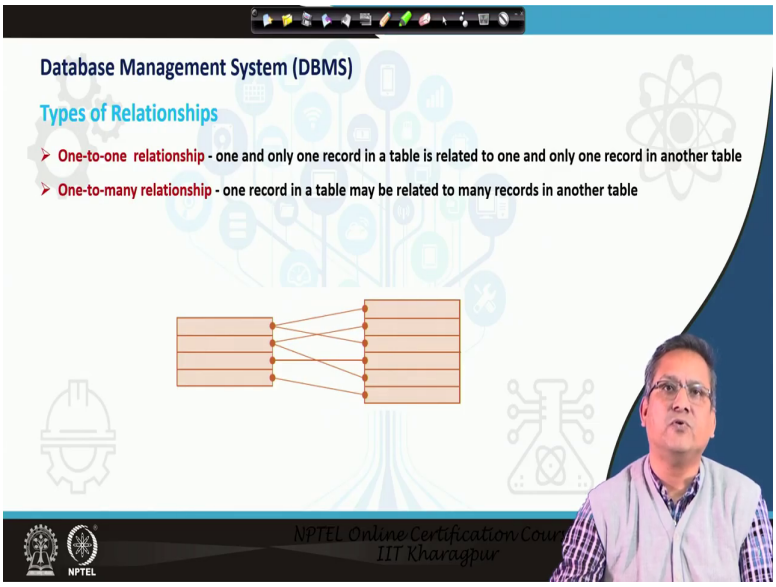
We had also talked about the BLOB which is the binary large object which stores of huge block of data I mean for example it could be the coordinates of the I mean the points or the lines I mean for I mean generally we store the feature geometries as block files. So, it could be also images or it could be multimedia data and these I mean files generally store this coordinates as binary numbers.

(Refer Slide Time: 28:55)



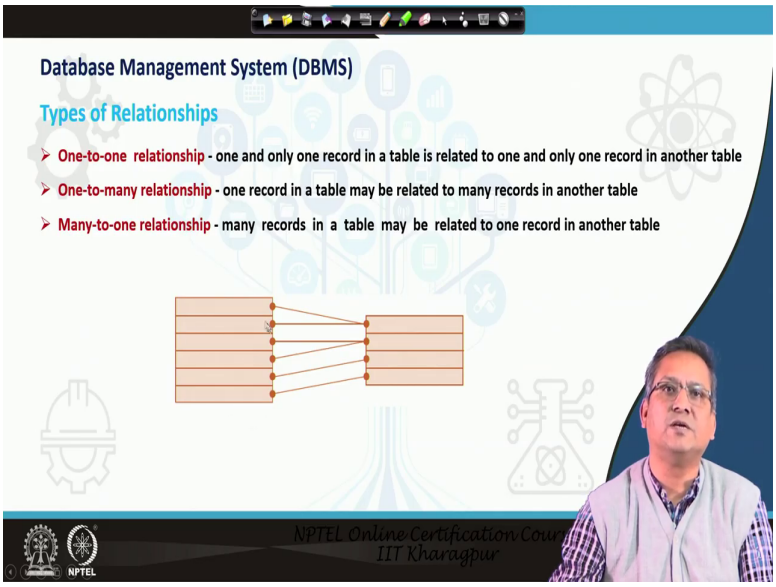
Now, the types of relationships that we have in the database management system could be of the four types. The first one is the one-to-one relationship. So, each record in the table is related to one and only record, one and only record in another table. So, if we have two tables it is a one-to-one relationships I mean you do not see multiple connections in this type of relationships.

(Refer Slide Time: 29:27)



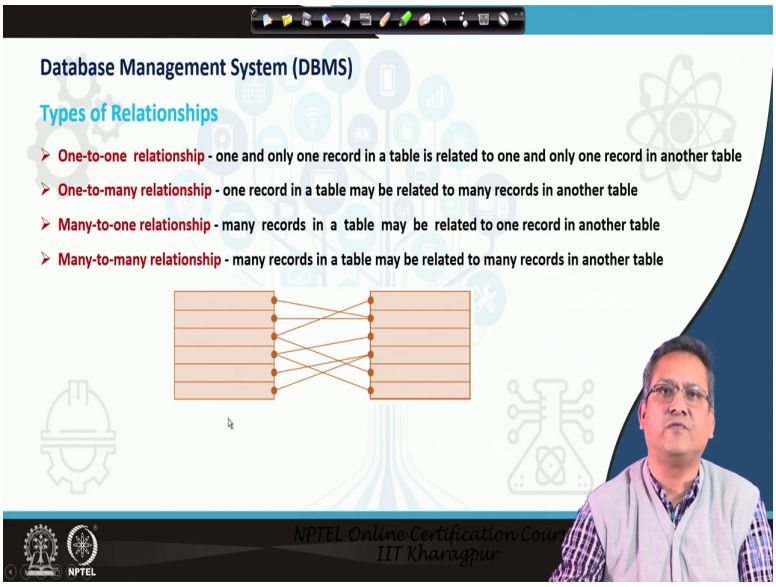
The next one that is the one-to-many relationship. In this one record of the table is related to many records in another table. So, you can see for each of these particular records they are related to multiple records in other table.

(Refer Slide Time: 29:44)



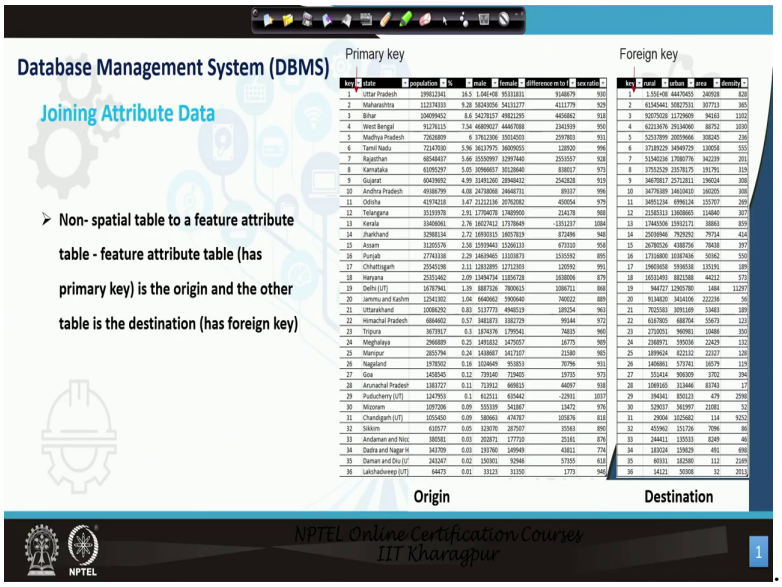
Now, there could be many-to-one relationship. In this many record in the table I mean your attribute table may be related to one record in another table. So, in this case you have your attributable table, GIS table and you may have another database where in you see there are multiple connections from your input database to your to another database another data table.

(Refer Slide Time: 30:13)



The next one is many-to-many relationships in which many records in a table would be related to many records in another table. So, this is how these two tables are. They would have their identifiers or the keys that we have talked about. So, I mean they would be related to each other in a many-to-many relationship.

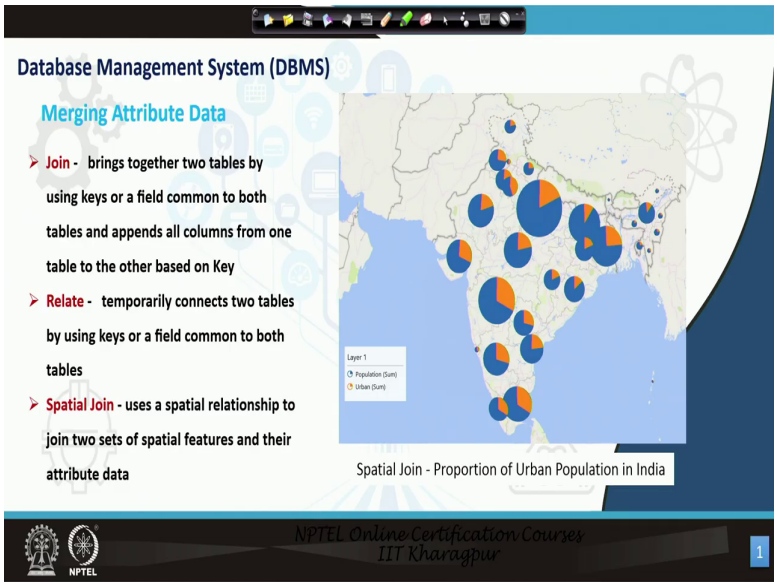
(Refer Slide Time: 30:34)



Now, talking about database management system we can join the attribute data I mean we can do a non-spatial join of the attribute table. So if we have multiple attribute table, in this case you can see that we have the population tables for different states of India in which we have population, we have the male and the female population, we have the difference between the male and the female and the sex ratio in the first table which has a primary key and in the next table, we have the total population which is urban and the total population which is rural we have the area as well as the density and it has a foreign key.

So, in this in the earlier slide we had talked about the primary key and the foreign key. So, you can see how they are located and they would be related or in any query or search operation or processing further processing these two keys would be related for in a non-spatial join. So, we can see how we can do a non-spatial join.

(Refer Slide Time: 31:55)

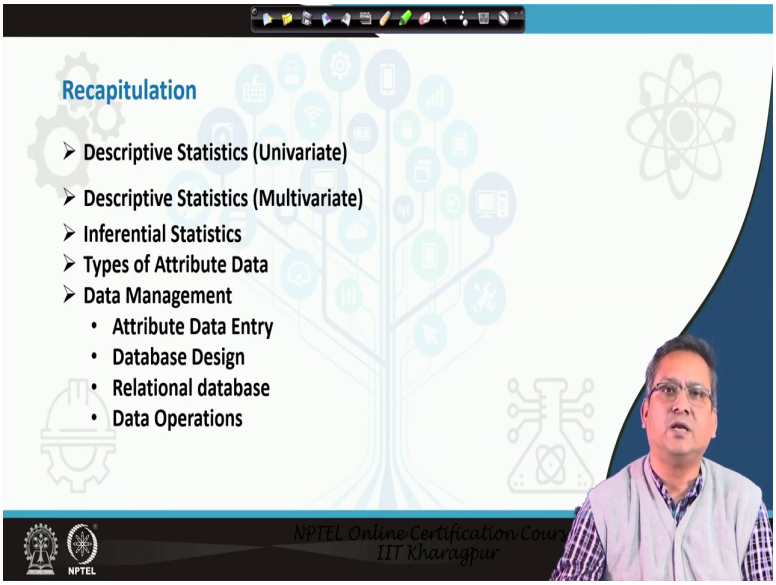


So, for merging the attribute data there are few options which are available in GIS softwares. So, most of these packages have got these operations. So, first one is the join operation where in the two tables using the keys, the common keys is are join and the columns are appended from one table to the another table. So, for the in the input table the I mean it would become an extended table, where in the all the fields would be appended.

The next one, next way to merge the attribute data is the relate operator, where in it temporally connects two tables by using the common keys or the fields that we have seen are common tool, both the tables. Now, the 3rd one is the spatial join which uses a spatial relationship to join the two data sets of the spatial features as well as their attribute data. So, we can see an example of spatial data, a spatial join.

So, in our earlier slide we had seen the data pertaining to the population statistics for different states of India. So, you can see here a spatial join operation has been done in the blue pi chart, pi you can see that it is the population and the orange one gives you the fraction of the urban population in the different states. So, you can identify the states where in which has the highest amount of urban population those two tables where join together using a spatial join and we can see the output here.

(Refer Slide Time: 33:54)



So, recapitulation of what we have covered today, we have talked about Univariate Descriptive Statistics, where in we had talked about the central tendencies of mean, median and mode for a univariate data which has only one attribute column. We had then talked about multivariate data analysis, we had talked about inferential statistics, we had talked about the different types of attribute data and then we have finally talked about data management.

We had talked about how we can do the data entry and what are the different types of data. In that we had talked about the database design, we had talked about the relational database, the data operations. So, thank you for your patient hearing till we meet again in the next lecture.

Thanks so much.