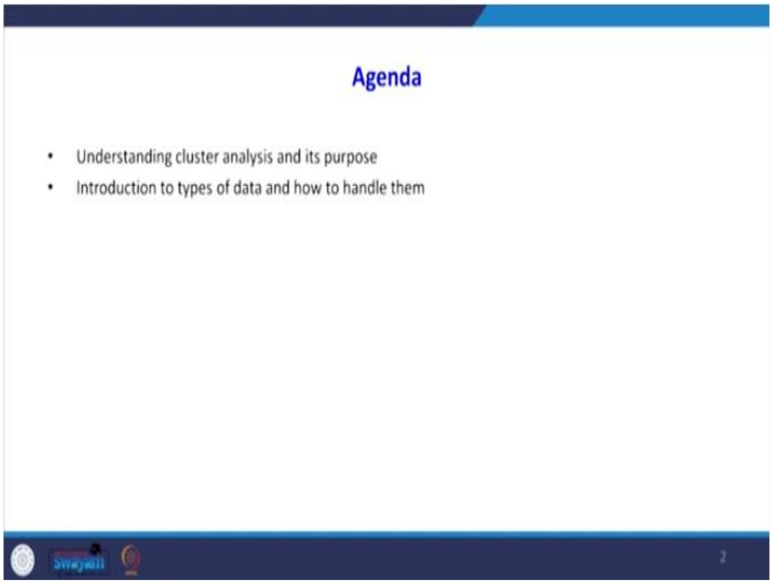
Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology - Roorkee

Lecture – 49
Cluster Analysis: Introduction - 1

(Refer Slide Time: 00:36)

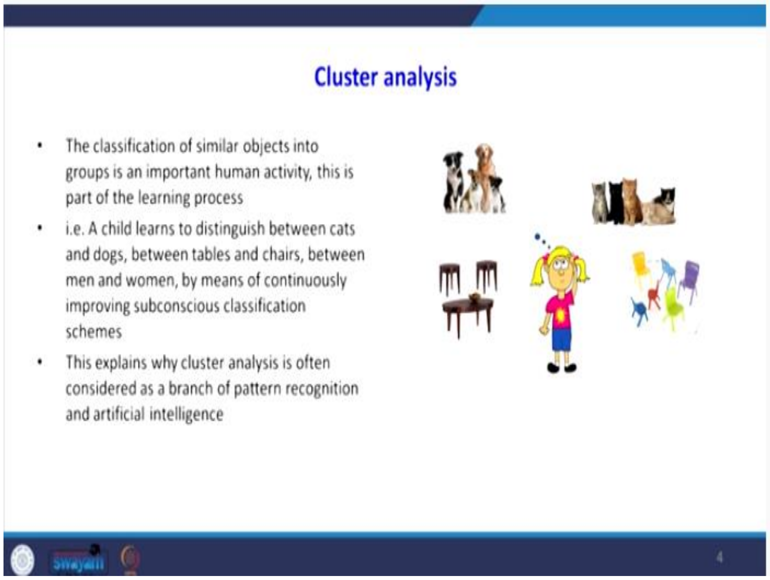


Agenda

- Understanding cluster analysis and its purpose
- Introduction to types of data and how to handle them

Dear students today we are entering to a new topic that is a cluster analysis. The cluster analysis is mostly widely used data mining techniques. It is a very important topic, so the agenda for today class is understanding cluster analysis and its purpose. Then introduction to types of data and how to handle them because the clustering techniques will vary with respect to what kind of data nature of the data is; suppose the nature of the data is continuous data or interval data, there will be a different algorithm for that. If the data is nominal data, there will be a different algorithm for clustering.

(Refer Slide Time: 01:06)



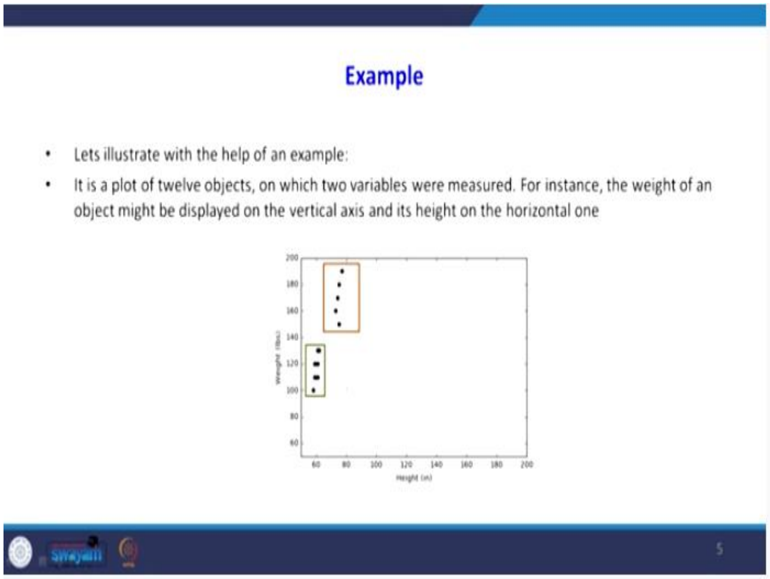
Cluster analysis

- The classification of similar objects into groups is an important human activity, this is part of the learning process
- i.e. A child learns to distinguish between cats and dogs, between tables and chairs, between men and women, by means of continuously improving subconscious classification schemes
- This explains why cluster analysis is often considered as a branch of pattern recognition and artificial intelligence

First, we will see what is cluster analysis? Cluster analysis is the art of finding groups in data. In cluster analysis basically one wants to form groups in such a way that objects in the same groups are similar to each other whereas the object in different groups are dissimilar as possible. When you look at this picture, they are different clusters say group of dogs, group of cat, group of chairs, group of tables. This is an example of cluster analysis.

The classification of similar objects into groups an important human activity. This is part of learning process. A child learns to distinguish between cats and dogs, between tables and chairs, between men and women by means of continuously improving subconscious classification schemes. What is the meaning of this point is a child unknowingly is able to classify different objects and able to group cluster different objects which are similar in nature.

(Refer Slide Time: 02:17)

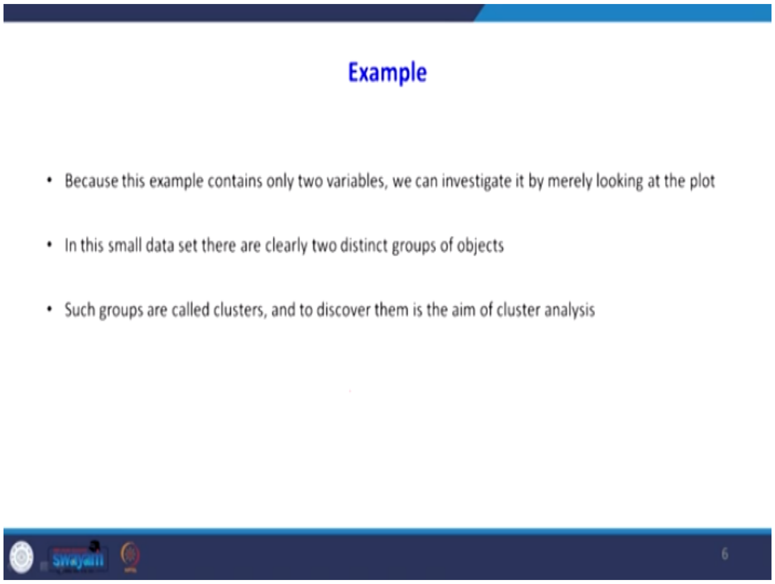


Example

- Lets illustrate with the help of an example:
- It is a plot of twelve objects, on which two variables were measured. For instance, the weight of an object might be displayed on the vertical axis and its height on the horizontal one

We will explain the concept of cluster analysis with the help of an example. This is a plot of 12 objects on which two variables were measured. For instance, the weight of an object might be displayed on the vertical axis and its height on the horizontal axis when you plot it here it is clearly visible that you are able to form two clusters with respect to height you see that these are the different weight with respect to another height to this is the 60 is one type of height and second is 80 is another type of height we are able to cluster it.

(Refer Slide Time: 02:56)

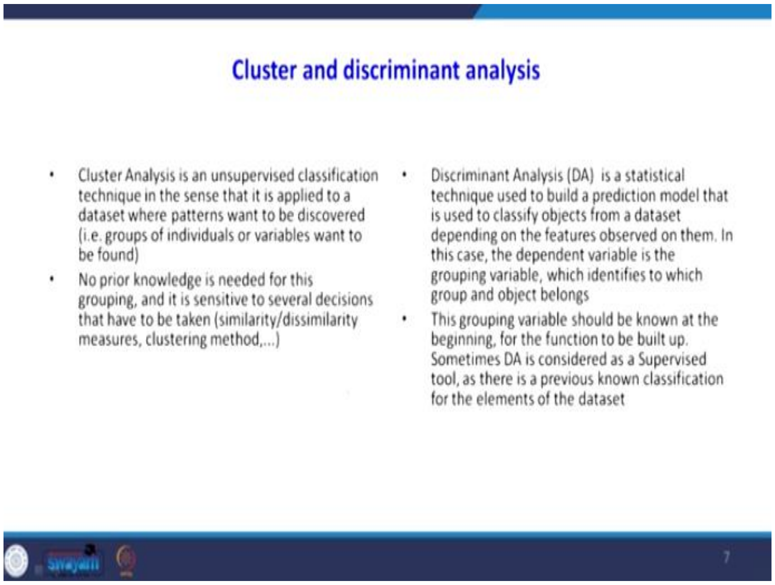


Example

- Because this example contains only two variables, we can investigate it by merely looking at the plot
- In this small data set there are clearly two distinct groups of objects
- Such groups are called clusters, and to discover them is the aim of cluster analysis

Because this is contains only two variable. We can investigate it by merely looking at the plot. In this small data set that are clearly two distinct group of objects. Such groups are called clusters and to discover them is the aim of cluster analysis, so this is the purpose of our class. What is going to be there in coming classes. We will be having different types of data. We are going to cluster that different data into different groups.

(Refer Slide Time: 03:32)



Cluster and discriminant analysis

- Cluster Analysis is an unsupervised classification technique in the sense that it is applied to a dataset where patterns want to be discovered (i.e. groups of individuals or variables want to be found)
- No prior knowledge is needed for this grouping, and it is sensitive to several decisions that have to be taken (similarity/dissimilarity measures, clustering method,...)

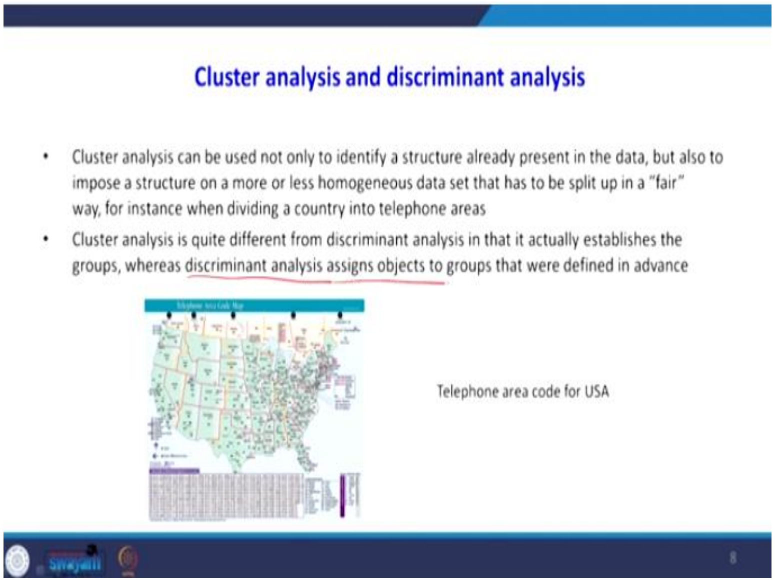
- Discriminant Analysis (DA) is a statistical technique used to build a prediction model that is used to classify objects from a dataset depending on the features observed on them. In this case, the dependent variable is the grouping variable, which identifies to which group and object belongs
- This grouping variable should be known at the beginning, for the function to be built up. Sometimes DA is considered as a Supervised tool, as there is a previous known classification for the elements of the dataset

Many time the students may have doubt what is the difference between cluster and discriminant analysis. Cluster analysis is an unsupervised classification technique in the sense that it is applied to a data set where patterns want to be discovered. That is the group of individuals or variables wanted to be found. Why we are calling it this unsupervised learning because we may not know which variable will go to which cluster, we are not knowing also that how many clusters we are going to form it.

The second point in a cluster analysis, no prior knowledge is needed for this grouping. I need to sensitive to several decisions that have to be taken. Some of the variables are similarity dissimilarity measures clustering methods. Whereas discriminant analysis is a statistical techniques used to build a prediction model that is used to classify objects from your data set depending on the futures observed on them.

In this case, the dependent variable is grouping variable which identifies to which group or object belongs. This grouping variable should be known at the beginning for a function to be built up sometime discriminant analysis is considers supervised tool because as there is a previous known classification for the element of the dataset.

(Refer Slide Time: 05:03)



Cluster analysis and discriminant analysis

- Cluster analysis can be used not only to identify a structure already present in the data, but also to impose a structure on a more or less homogeneous data set that has to be split up in a "fair" way, for instance when dividing a country into telephone areas
- Cluster analysis is quite different from discriminant analysis in that it actually establishes the groups, whereas discriminant analysis assigns objects to groups that were defined in advance

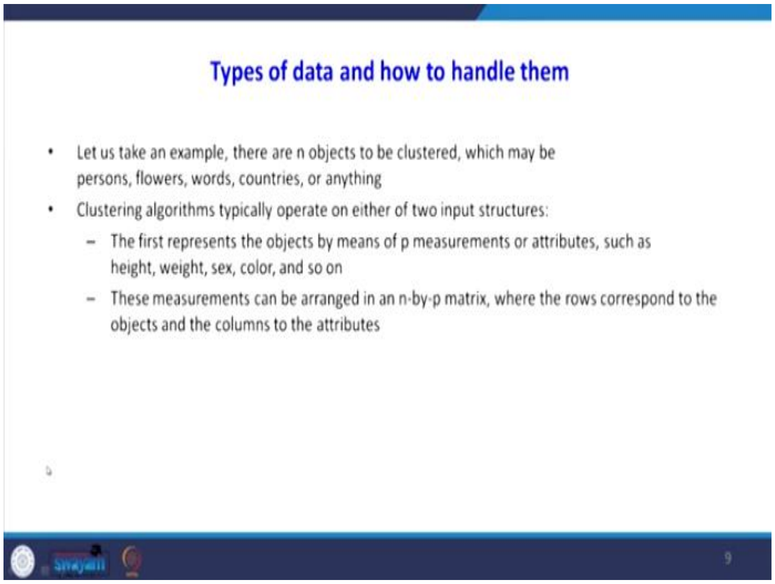
Telephone area code for USA

8

Further we will continue the difference between a cluster analysis and discriminant analysis. Cluster analysis can be used not only to identify the structure already present in the data, but also to impose structure on a more or less homogeneous data set that has to be split up in a fair way. For instance, when dividing a country into telephone areas. See this is a country for example see this example this is classified into different telephone areas.

Cluster analysis is quite different from discriminant analysis in that it actually establishes the groups whereas discriminant analysis assigns object to groups that were defined in advance. That is a major difference. What does that mean? The discriminant analysis assigns object to the group that were defined in advance. But in cluster analysis it is not the case and what will happen as I told you in the beginning, the clustering analysis and corresponding algorithm depending upon what kind of data.

(Refer Slide Time: 06:07)



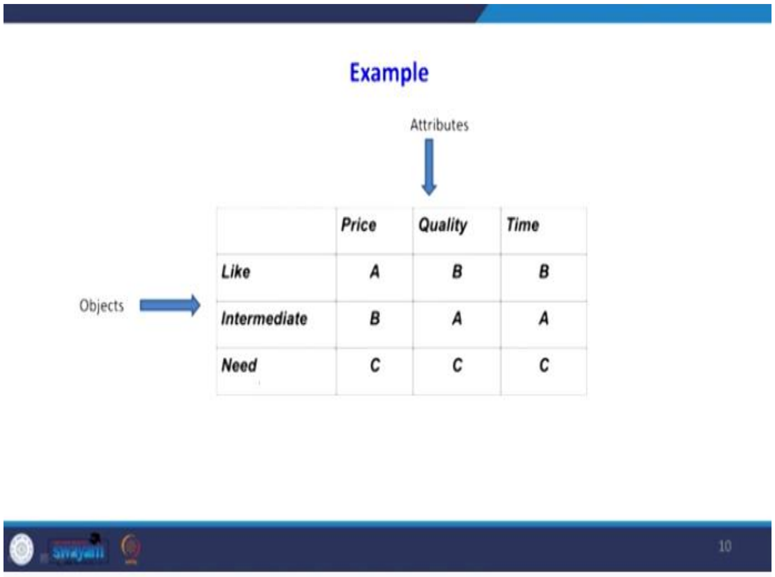
Types of data and how to handle them

- Let us take an example, there are n objects to be clustered, which may be persons, flowers, words, countries, or anything
- Clustering algorithms typically operate on either of two input structures:
  - The first represents the objects by means of p measurements or attributes, such as height, weight, sex, color, and so on
  - These measurements can be arranged in an n-by-p matrix, where the rows correspond to the objects and the columns to the attributes

Types of data and how to handle them for cluster analysis, as I told you in the beginning of the class, the types of data is an important point has to be taken care while doing cluster analysis because for different types of data there is a different type of clustering algorithms. Let us take an example that are n objects to be clustered, which may be persons, flowers, birds, countries or anything.

Clustering algorithm typically operate on either of two input structure. The first represents the object by means of p measurement or attributes such as height, weight, sex color and so on. These measurements can be arranged in a n-by-p matrix whereas the row corresponds to the objects and the column corresponds to the attributes.

(Refer Slide Time: 07:01)



Example

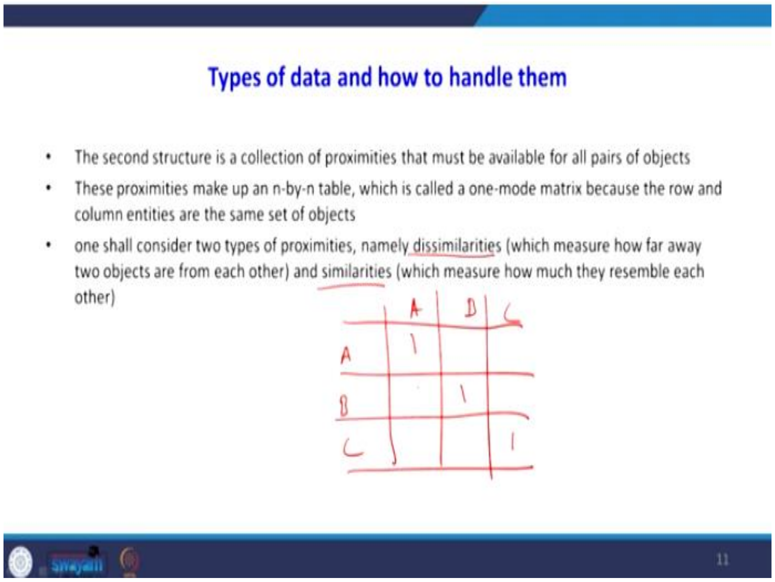
Attributes

Objects

	Price	Quality	Time
Like	A	B	B
Intermediate	B	A	A
Need	C	C	C

You see this case the objects are in the rows Like, Intermediate, Need the attributes, Price, Quality, Times are in the columns this is one kind of input.

(Refer Slide Time: 07:11)



Types of data and how to handle them

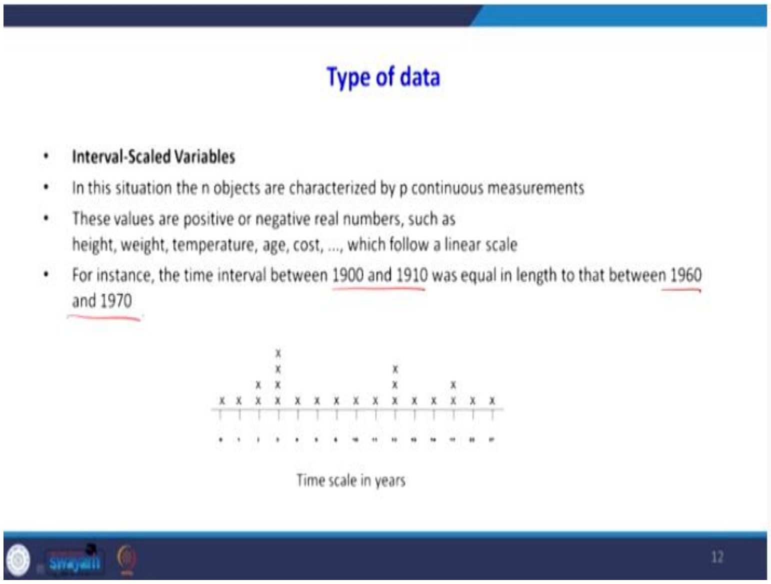
- The second structure is a collection of proximities that must be available for all pairs of objects
- These proximities make up an n-by-n table, which is called a one-mode matrix because the row and column entities are the same set of objects
- one shall consider two types of proximities, namely dissimilarities (which measure how far away two objects are from each other) and similarities (which measure how much they resemble each other)

	A	B	C
A	1		
B		1	
C			1

The second structure is a collection of proximities that must be available for all pairs of objects. These proximities makeup an n by n table, which is called one mode matrix because the row and the column entities are the same set of objects. One shall consider two type of proximities, namely dissimilarities which measures how far away two objects are from each other and similarities which measures how much they are resemble each other okay.

Now assume that there are some variable A, B, C, so the A and B it is written this way. You see that A and A, B and B, C and C is 1 because the same one. So we can write between A and B how much is the similarity otherwise between A and B how much is dissimilarity.

(Refer Slide Time: 08:18)



Type of data

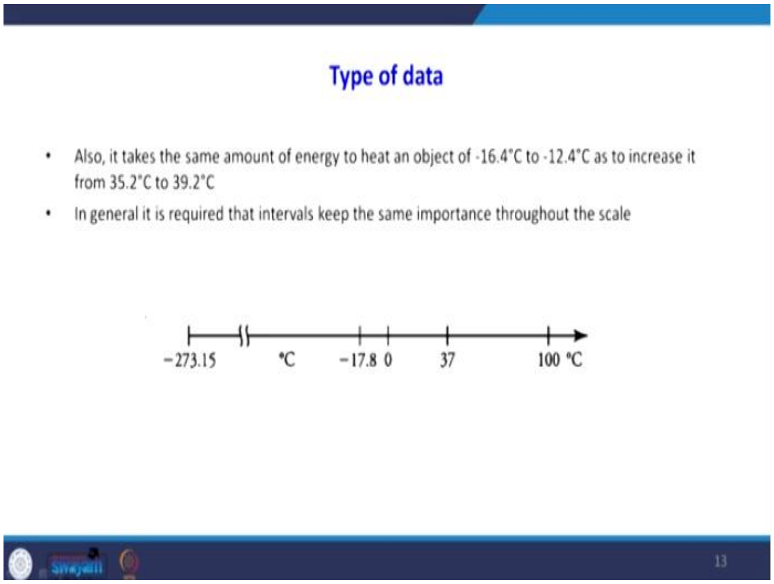
- **Interval-Scaled Variables**
- In this situation the n objects are characterized by p continuous measurements
- These values are positive or negative real numbers, such as height, weight, temperature, age, cost, ..., which follow a linear scale
- For instance, the time interval between 1900 and 1910 was equal in length to that between 1960 and 1970

12

Now let us see what is the interval scaled variables. In this situation the n objects are characterized by p continuous measurement. These values are positive or negative real numbers such as height, weight, temperature, age, cost which follow a linear scale. For instance, the time interval between 1900 and 1910 was equal in length to that between 1960 and 1970. So this is an example of interval scales we have studied in the beginning of the lecture.

We have classified data in to different four categories, nominal, ordinal, interval, ratio. So, the example of interval is year. If it is a year, what happened? We can add some numbers, we can subtract some numbers. Similarly, you see that between 19 the interval will be same between 1900 to 1910 and 1960 to 1970 the difference is same because we can add it, we can subtract it, but we cannot multiply.

(Refer Slide Time: 09:18)

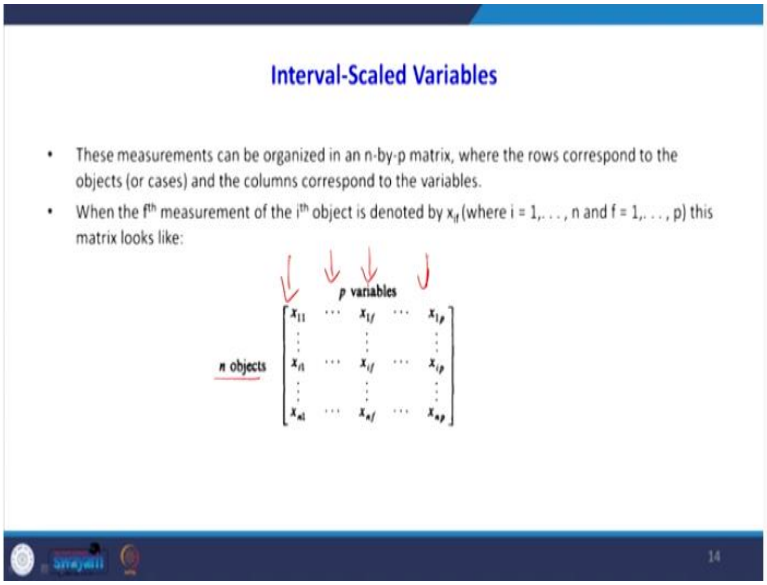


Type of data

- Also, it takes the same amount of energy to heat an object of -16.4°C to -12.4°C as to increase it from 35.2°C to 39.2°C
- In general it is required that intervals keep the same importance throughout the scale

Similarly the another example for interval data is our temperature. Also it takes the same amount of energy to heat an object of - 14.4 degrees Celsius to - 12.4 degrees Celsius as to increase it from 35.2 degrees Celsius to 39.2 degree Celsius. What I am saying is that the Fahrenheit temperature scale also an example of interval scale because there would not be any absolute 0 but we can add, we can subtract. In general it is required that intervals keeps the same importance throughout the scale.

(Refer Slide Time: 09:55)



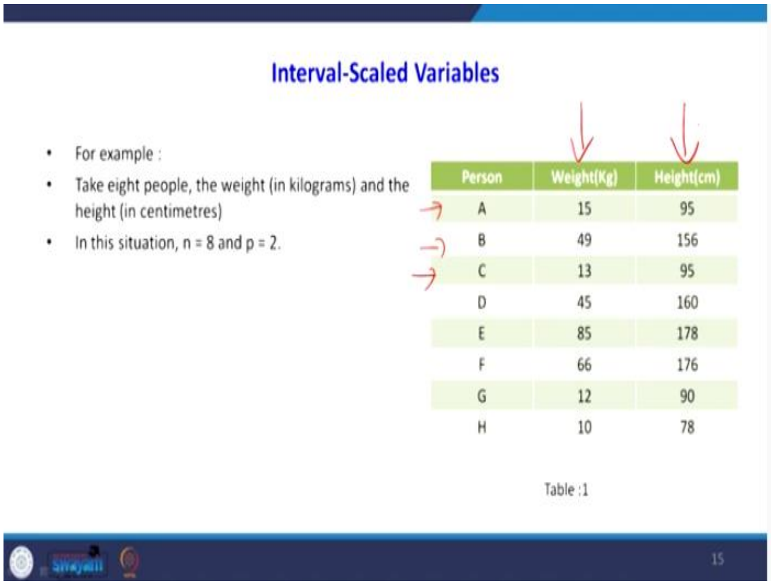
Interval-Scaled Variables

- These measurements can be organized in an n-by-p matrix, where the rows correspond to the objects (or cases) and the columns correspond to the variables.
- When the fth measurement of the ith object is denoted by x_{if} (where i = 1,. . . , n and f = 1,. . . , p) this matrix looks like:

$$\begin{array}{cc} & p \text{ variables} \\ n \text{ objects} & \begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix} \end{array}$$

Interval scale scaled variables. These measurements can be organized in an n-by-p matrix where the rows corresponds to the objects and the column corresponds to the variables. So where, the fth measurement of the ith object is denoted by xif where i is 1 to n , f is 1 to p . So, here in row we have mentioned objects in column we have mentioned variables. The another name for object is cases the variable may be different variables

(Refer Slide Time: 10:33)



Interval-Scaled Variables

- For example :
- Take eight people, the weight (in kilograms) and the height (in centimetres)
- In this situation, n = 8 and p = 2.

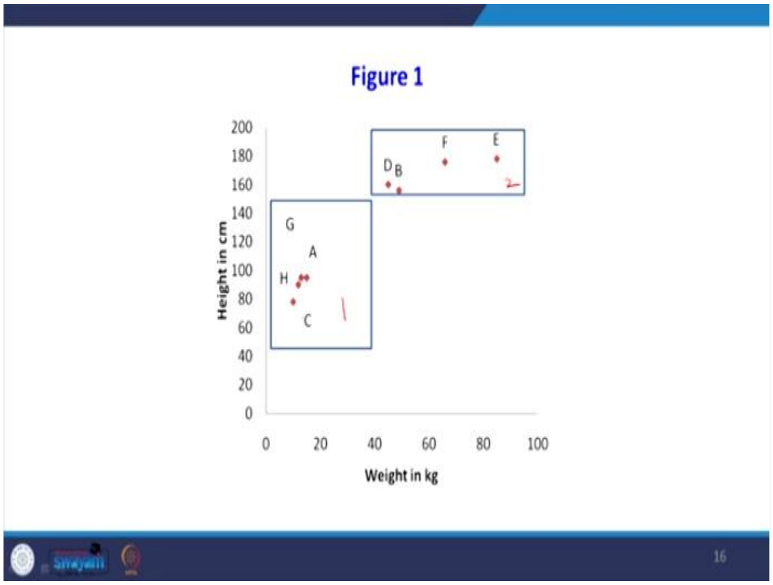
Person	Weight(Kg)	Height(cm)
A	15	95
B	49	156
C	13	95
D	45	160
E	85	178
F	66	176
G	12	90
H	10	78

Table :1

For example take 8 people for example what is happening in rows we have n objects in column we have 2 variables one variable is height another variable is weight. Take 8 people the weight in kilogram and the height in centimeter is given in the table. In this situation n = 8 because 8 rows are there p =2 because 2 variable is there.

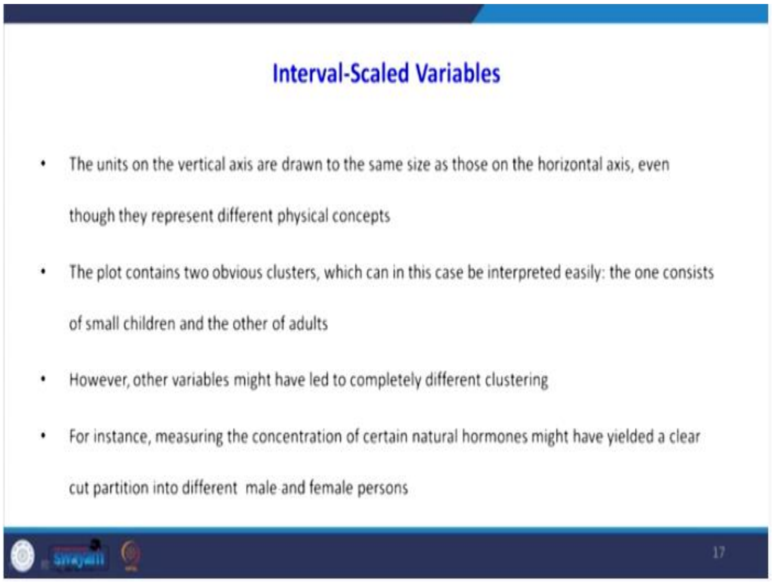
(Refer Slide Time: 11:01)

Figure 1



If I plot that you see that weight is taken in kg height in centimeter. When you plot it, it is forming two similar objects. We can group into similar objects into two category one is cluster 1 we can call it is cluster 1 and cluster 2. In cluster 1 for example your row C H A G will occur in cluster 2 row D B F E will occur. This is a one way of clustering.

(Refer Slide Time: 11:35)



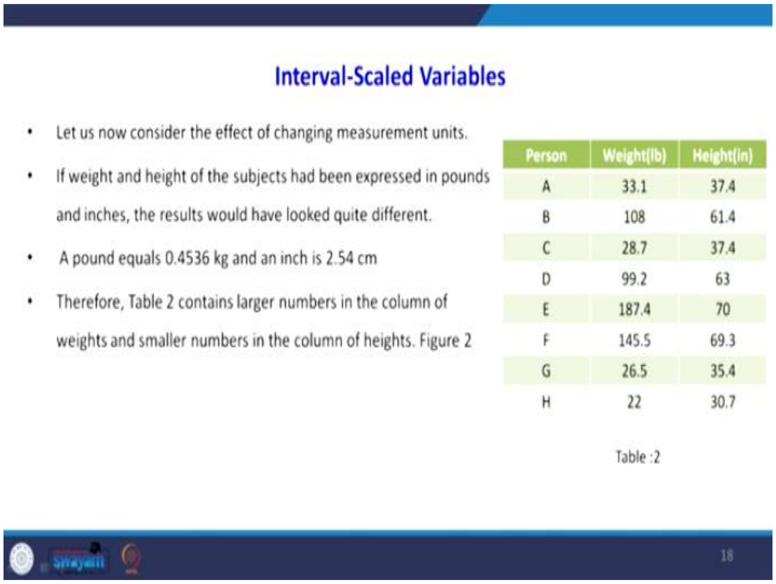
Interval-Scaled Variables

- The units on the vertical axis are drawn to the same size as those on the horizontal axis, even though they represent different physical concepts
- The plot contains two obvious clusters, which can in this case be interpreted easily: the one consists of small children and the other of adults
- However, other variables might have led to completely different clustering
- For instance, measuring the concentration of certain natural hormones might have yielded a clear cut partition into different male and female persons

17

The units on the vertical axis are drawn to the same size as those under horizontal axis even though there represents different physical concepts. The plot contains two obvious clusters which can in this case be interpreted easily. The one consist of small children other of adult. However other variables might have led to completely different clustering. For instance, measuring the concentration of certain natural hormones might have yielded a clear cut partition into different male and female persons. In this one since we are taken 2 variable one is weight and height instead of weight and height if you take some other variables that may bring some other type of clustering.

(Refer Slide Time: 12:21)



Interval-Scaled Variables

- Let us now consider the effect of changing measurement units.
- If weight and height of the subjects had been expressed in pounds and inches, the results would have looked quite different.
- A pound equals 0.4536 kg and an inch is 2.54 cm
- Therefore, Table 2 contains larger numbers in the column of weights and smaller numbers in the column of heights. Figure 2

Person	Weight(lb)	Height(in)
A	33.1	37.4
B	108	61.4
C	28.7	37.4
D	99.2	63
E	187.4	70
F	145.5	69.3
G	26.5	35.4
H	22	30.7

Table :2

Let us now consider the effect of changing the measurement unit. So previously we had the measurement unit and you look at this one the weight is in kg and height is in centimeter. Now if you change the unit now, the weights going to be in pound and height is going to be in inch. Now for this kind of dataset, let us see how this unit of this data is going to affect our clustering technique.

Let us now consider the effect of changing measurement units. If weight and height of the subject had been expressed in pounds and inches the result would have looked quite different. A pound equals 0.453 kg and an inch is 2.5 centimeter. Therefore, table 2 contains larger number in column of weight because we are converted into pounds and smaller number in the column of heights the heights become very smaller.

(Refer Slide Time: 13:18)

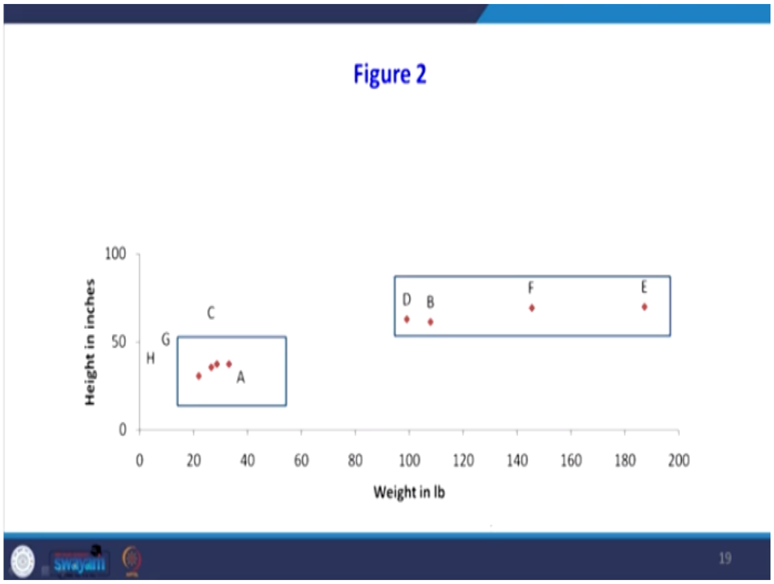
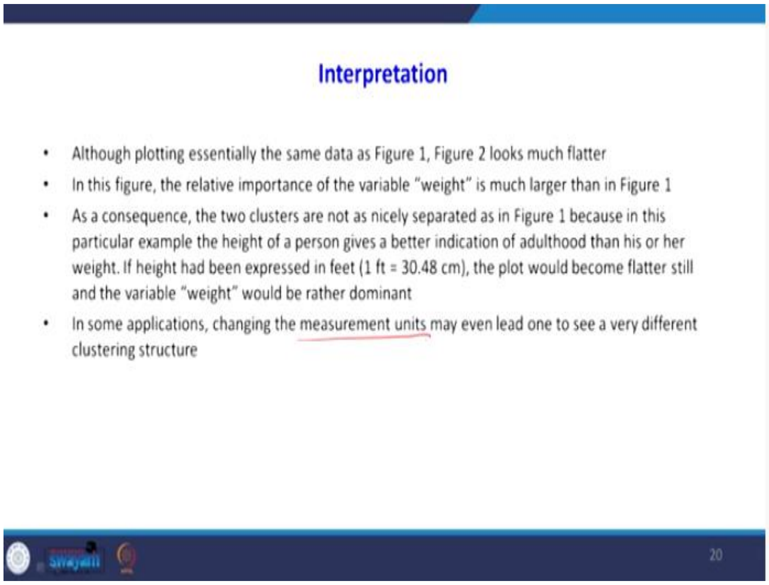


Figure 2

Let us see the new cluster now what happened? Now the height has increased, the weight has increased. Now the clustering pattern is completely changed. So what point I am saying here is that the unit of the data may bring out different clusters.

(Refer Slide Time: 13:36)



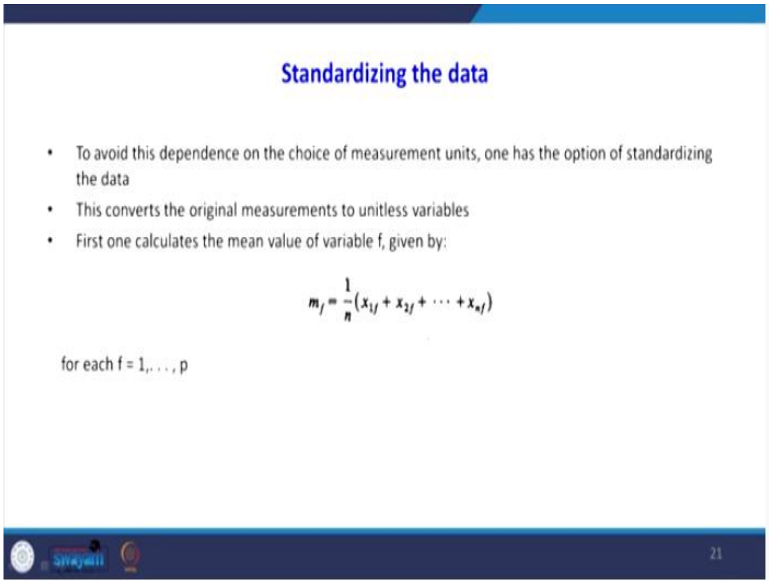
Interpretation

- Although plotting essentially the same data as Figure 1, Figure 2 looks much flatter
- In this figure, the relative importance of the variable "weight" is much larger than in Figure 1
- As a consequence, the two clusters are not as nicely separated as in Figure 1 because in this particular example the height of a person gives a better indication of adulthood than his or her weight. If height had been expressed in feet (1 ft = 30.48 cm), the plot would become flatter still and the variable "weight" would be rather dominant
- In some applications, changing the measurement units may even lead one to see a very different clustering structure

So what is interpretation, although plotting essentially the same data as figure 1, figure 2 looks much flatter. In this figure the relative importance of the variable weight is much larger than the figure 1. As a consequence, the two clusters are not as nicely separated as in figure 1 because in this particular example, the height of the person gives a better indication of adulthood then his or her weight.

If height had been expressed in feet because 1 feet = 30.48 centimeter the plot would become flatter still and the variable weight would be rather dominant. In some applications changing the measurement units that is an important point, may even lead to one to see a very different clustering structures. The point what I am trying to say here is that changing the measurement unit may provide different type of clustering structure.

(Refer Slide Time: 14:37)



Standardizing the data

- To avoid this dependence on the choice of measurement units, one has the option of standardizing the data
- This converts the original measurements to unitless variables
- First one calculates the mean value of variable f, given by:

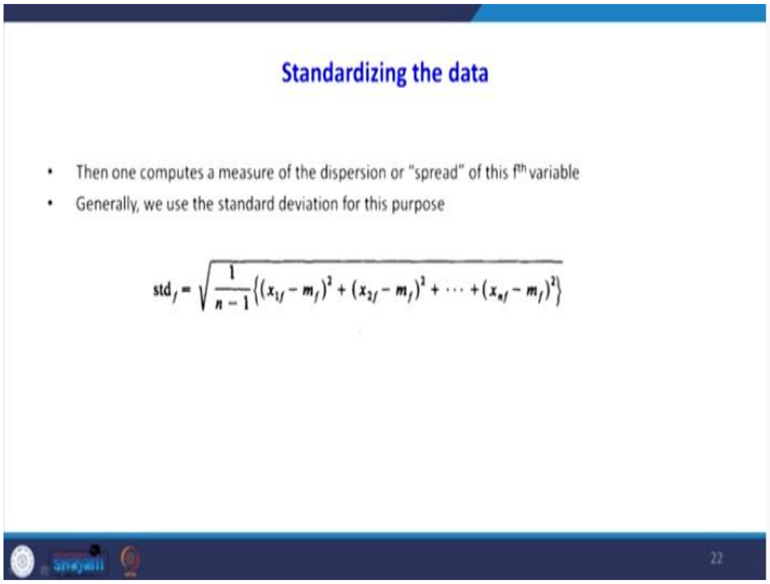
$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \cdots + x_{nf})$$

for each f = 1,. . . , p

21

To avoid this because different units are providing different clustering structure one way to avoid this problem is standardizing the data. Now let us see how to standardize the data to avoid this dependence on the choice of measurement units one has the option of standardizing the data. This converts the original measurement to unitless variable. First one calculates the mean value of f given by we know that the mean is 1 / n sum of all the values divided by m number of variables that is mf.

(Refer Slide Time: 15:18)



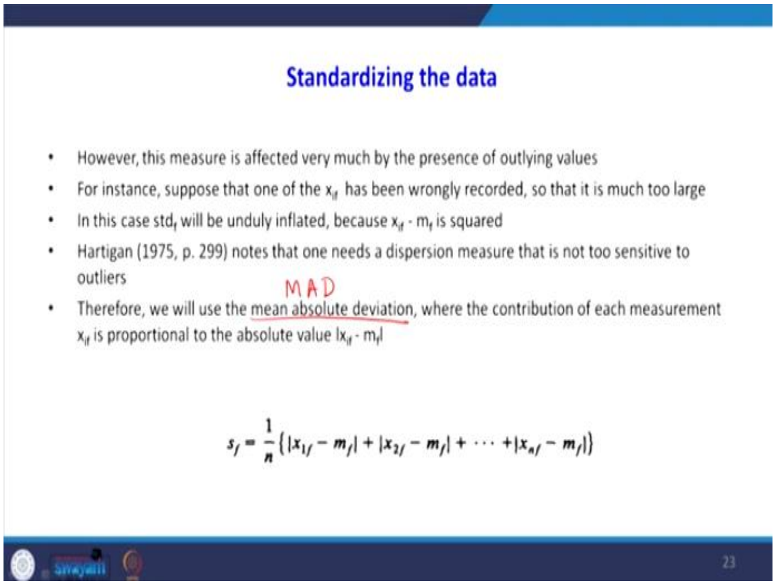
Standardizing the data

- Then one computes a measure of the dispersion or "spread" of this f^{th} variable
- Generally, we use the standard deviation for this purpose

$$\text{std}_f = \sqrt{\frac{1}{n-1}\left\{(x_{1f}-m_f)^2 + (x_{2f}-m_f)^2 + \cdots + (x_{nf}-m_f)^2\right\}}$$

Then one computes a measure of dispersion or spread of this fth variable. Generally we use the standard deviation what is a standard deviation x1 that variable minus mean whole square divided by second variable minus whole square up to fth variable - mf whole square divided by n – 1 that is a standard deviation. This is one way of standardizing the data.

(Refer Slide Time: 15:44)



Standardizing the data

- However, this measure is affected very much by the presence of outlying values
- For instance, suppose that one of the x_{if} has been wrongly recorded, so that it is much too large
- In this case std_f will be unduly inflated, because $x_{if} - m_f$ is squared
- Hartigan (1975, p. 299) notes that one needs a dispersion measure that is not too sensitive to outliers
- Therefore, we will use the mean absolute deviation (MAD), where the contribution of each measurement x_{if} is proportional to the absolute value $|x_{if} - m_f|$

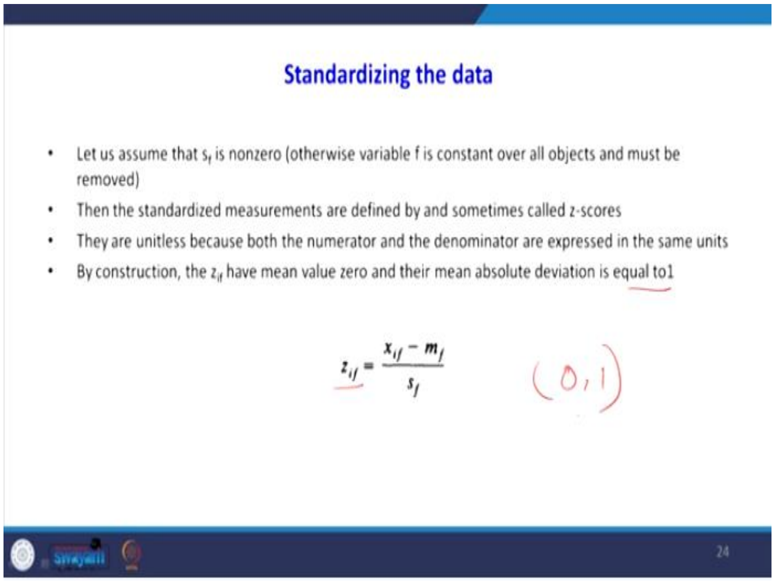
$$s_f = \frac{1}{n}\left\{|x_{1f}-m_f| + |x_{2f}-m_f| + \cdots + |x_{nf}-m_f|\right\}$$

However this measure is affected very much by the presence of outlying values. The problem with the standardization is that if there are extremely large values or extremely low values that is affecting the process of standardization. For instance, suppose that one of the xif has been wrongly recorded so that it is much too large. In this case, the standard deviation will be unduly inflated because we are squaring xif – m is squared .

So Hartigan in the year 1975 notes that one needs a dispersion measure that is not too sensitive to outliers. Therefore we will use the mean absolute deviation we generally this term generally we call it as MAD mean absolute deviation where the contribution of each measurement xif is proportional to the absolute value of modulus value of xif – mf. So instead of squaring, we are going to take to find the standardize we are going to take only the mean absolute deviation.

The advantage of taking mean absolute deviation is that if any out layer is there that its effect is dampened. That is why instead of going for standard deviation, we should go for mean absolute deviation. That is xf = 1/ n modulus of x1f - mf + x2f – mf modulus and so on.

(Refer Slide Time: 17:21)



Standardizing the data

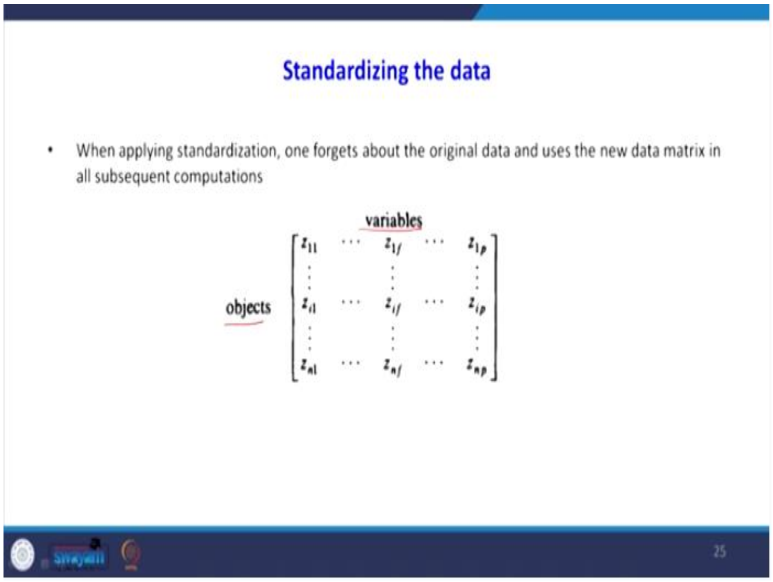
- Let us assume that s_f is nonzero (otherwise variable f is constant over all objects and must be removed)
- Then the standardized measurements are defined by and sometimes called z-scores
- They are unitless because both the numerator and the denominator are expressed in the same units
- By construction, the z_{if} have mean value zero and their mean absolute deviation is equal to1

$$\underline{z_{if}} = \frac{x_{if} - m_f}{s_f} \qquad (0,1)$$

Let us assume that Sf is nonzero because the standardized value should be non-zero otherwise the variable f is constant over all objects and must be removed. Then the standardized measurement are defined by and sometimes called z-scores. The another name for standardization is z-score. They are unitless because both the numerator, because numerator also deviation the denominators are also deviation.

They are unitless because the numerator, the denominator are expressed in the same units. By construction zif have mean 0 and then absolute deviation is equal to 1. So what is happening in the property of this where standardized data is mean should be 0 and the variance or standard division should be 1.

(Refer Slide Time: 18:10)



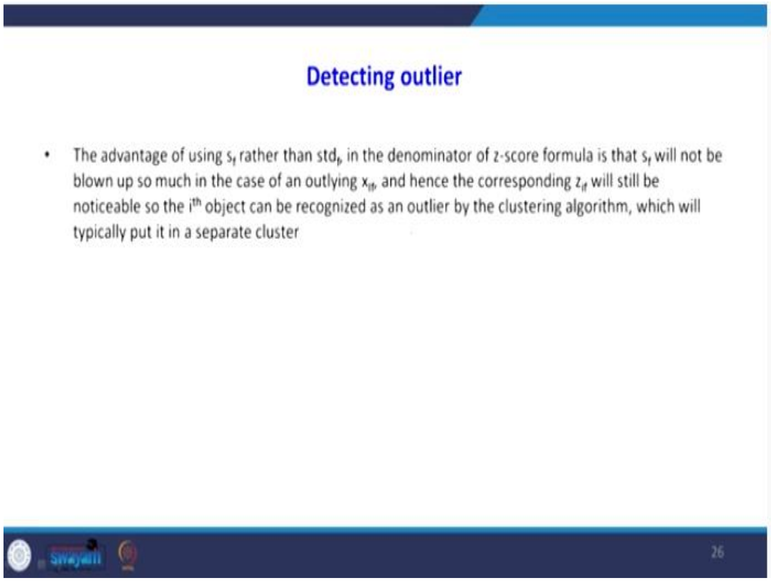
Standardizing the data

- When applying standardization, one forgets about the original data and uses the new data matrix in all subsequent computations

$$\begin{array}{cc} & \text{variables} \\ \text{objects} & \begin{bmatrix} z_{11} & \cdots & z_{1f} & \cdots & z_{1p} \\ \vdots & & \vdots & & \vdots \\ z_{i1} & \cdots & z_{if} & \cdots & z_{ip} \\ \vdots & & \vdots & & \vdots \\ z_{n1} & \cdots & z_{nf} & \cdots & z_{np} \end{bmatrix} \end{array}$$

When applying standardization one forgets about the original data and uses the new data matrix in all subsequent computations. What happened the initially in row there was object in column there was variables. There was xif variable was there. So after standardization, that will become z11 z12 up to z1p. Now this data that is data which are standardized data will be taken for further analysis.

(Refer Slide Time: 18:40)

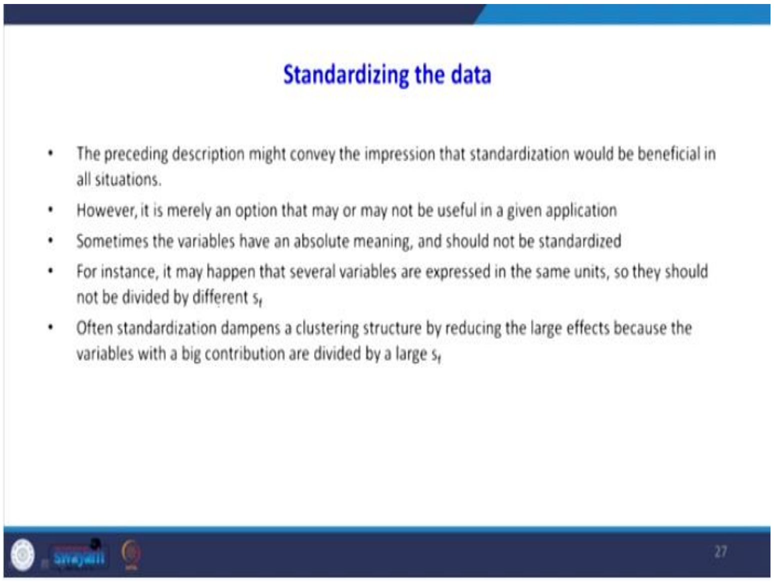


Detecting outlier

- The advantage of using s_f rather than std$_f$, in the denominator of z-score formula is that s_f will not be blown up so much in the case of an outlying x_{if}, and hence the corresponding z_{if} will still be noticeable so the ith object can be recognized as an outlier by the clustering algorithm, which will typically put it in a separate cluster

Detecting outlier the advantage of using Sf rather than standardized value of f in the denominator of z-score formula is that Sf will not be blown up so much in the case of an outlying xif and hence the corresponding zif will still be noticeable. So the ith object can be recognized as an outlier by the clustering algorithm, which will typically put in a separate cluster. So the purpose of using the z-score is it will not blown up so much in the case of any outlier in the dataset.

(Refer Slide Time: 19:18)



Standardizing the data

- The preceding description might convey the impression that standardization would be beneficial in all situations.
- However, it is merely an option that may or may not be useful in a given application
- Sometimes the variables have an absolute meaning, and should not be standardized
- For instance, it may happen that several variables are expressed in the same units, so they should not be divided by different s_f
- Often standardization dampens a clustering structure by reducing the large effects because the variables with a big contribution are divided by a large s_f

Standardizing the data the preceding description might convey the impression that the standardization would be beneficial in all situations. However, it is merely an option that may or may not be useful in a given application. Sometimes the variables have an absolute meaning and should not be standardized. What is the point here is that the variable already in the absolute term, it should not be standardized.

For instance, it may happen that several variables are expressed in the same units, so they should not be divided by different Sf. Because all the variables are in the same units you need not go for standardization. Often standardization dampens a clustering structure by reducing the large effect because the variables with the big contribution are divided by a large Sf. Sf is standardized value.

In this lecture we have covered the purpose of clustering analysis. Then I have explained the difference between clustering analysis and discriminant analysis. Then I have explained how the different types of data will affect our clustering structure. In the different types of data we have taken only the interval data and how to handle them for doing cluster analysis.

Then I have started why we have to do the standardization because if the different variables are in different units, you may get different kinds of clustering structures. To overcome that we have to go for standardization. The next class I will explain that the standardization also not applicable for all kind of data. Sometime it will mislead it may provide a different type of clustering that I will explain with the help of an example in the next class. Thank you.