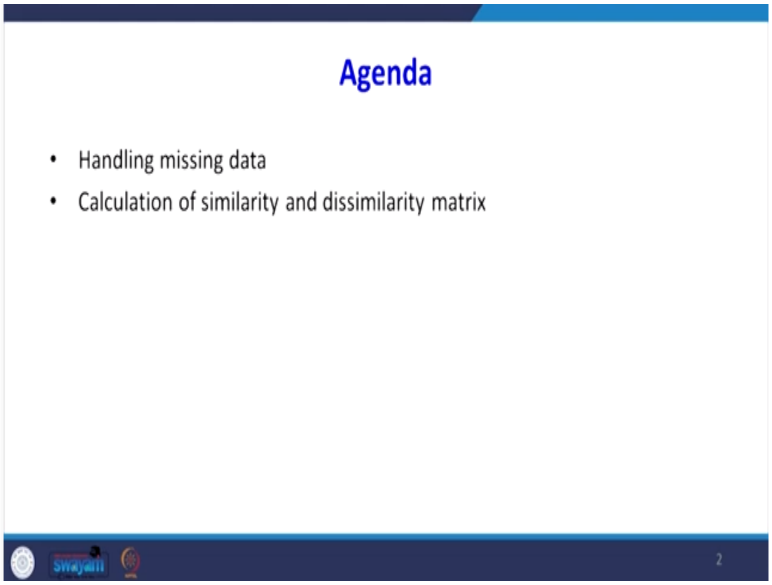
Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology - Roorkee

Lecture – 51
Clustering analysis: Part III

(Refer Slide Time: 00:46)

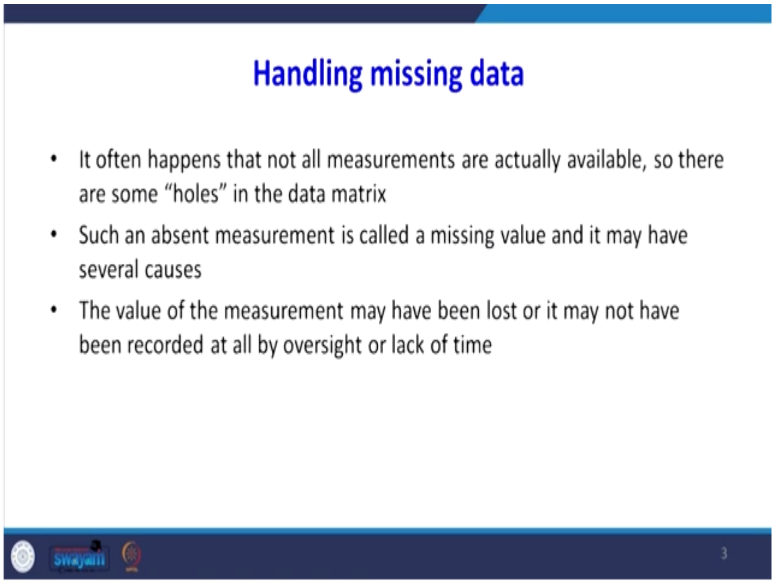


Agenda

- Handling missing data
- Calculation of similarity and dissimilarity matrix

In our previous class we have seen effect of standardization and how to find out different distances like Manhattan distance, Euclidean distance and Minkowski distance. Then I have explained how to select the variables. In this lecture we are going to see when you are collecting the data if you are missing some data, some data is not available how to handle that situation. Then very important concept of similarity and dissimilarity matrix. That is our agenda for this lecture.

(Refer Slide Time: 01:00)



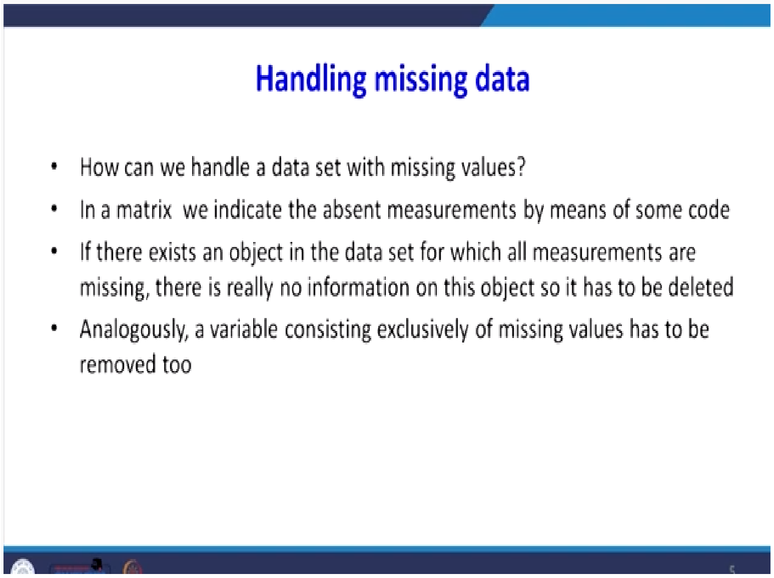
Handling missing data

- It often happens that not all measurements are actually available, so there are some “holes” in the data matrix
- Such an absent measurement is called a missing value and it may have several causes
- The value of the measurement may have been lost or it may not have been recorded at all by oversight or lack of time

First let us see how to handle the missing data. It is often happens that not all measurements are actually available. So there are some holes in the data matrix that is a missing value in the data matrix. Such an absent measurement is called missing value it may have several causes. The value of measurement may have been lost or it may not have been recorded at all by oversight or lack of time.

Sometime the information is simply not available. For example, birth date of a foundling or the patients may not remember whether he or she ever had their measles, or it may be impossible to measure the desired quantity due to the malfunctioning of some instrument. In certain instances, the question does not apply or there may be more than one possible answer when the experiments obtain very different results.

(Refer Slide Time: 02:07)

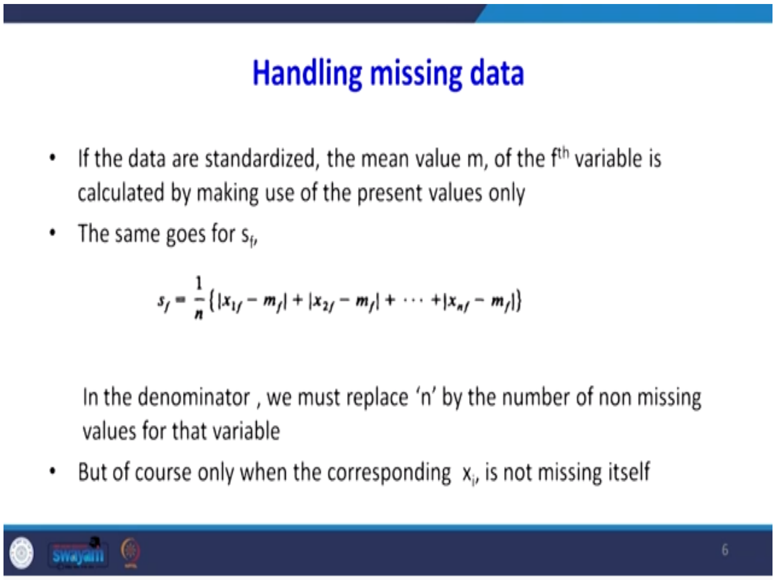


Handling missing data

- How can we handle a data set with missing values?
- In a matrix we indicate the absent measurements by means of some code
- If there exists an object in the data set for which all measurements are missing, there is really no information on this object so it has to be deleted
- Analogously, a variable consisting exclusively of missing values has to be removed too

So how can we handle a data set with the missing values? That is important question now in a matrix we indicate that the absent measurement by means of some code. If there exists an object in the dataset for which all measurements are missing, there is really no information on this object, so it has to be deleted. Analogously a variable consisting exclusively of missing values has to be removed too.

(Refer Slide Time: 02:37)



Handling missing data

- If the data are standardized, the mean value m, of the f^{th} variable is calculated by making use of the present values only
- The same goes for s_f,

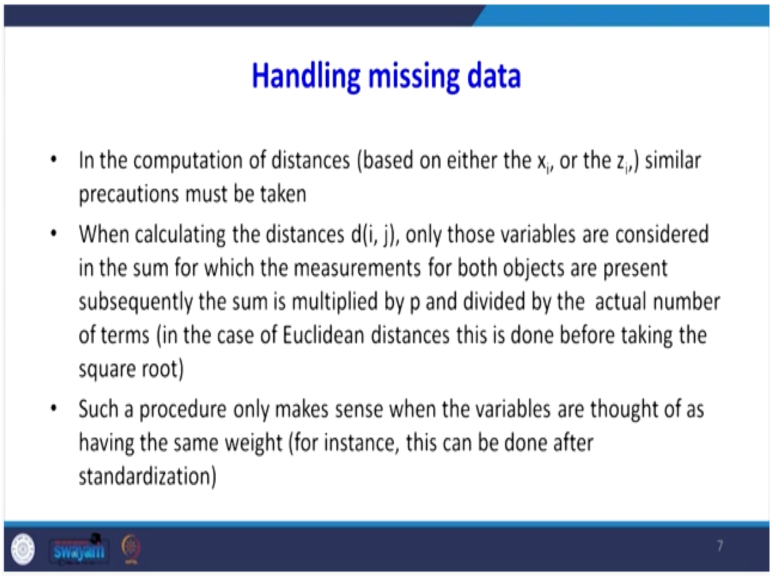
$$s_f = \frac{1}{n}\{|x_{1f} - m_f| + |x_{2f} - m_f| + \cdots + |x_{nf} - m_f|\}$$

In the denominator , we must replace 'n' by the number of non missing values for that variable

- But of course only when the corresponding x_i, is not missing itself

If the data are standardized, the mean value m of the fth variable is calculated by making use of present values only. The same goes for your mean absolute deviation, so mean absolute deviation that is Sf = (1 / n) modulus of (x1f – mf) and so on l xnf –mf l. In the denominator, we must replace n by the number of non-missing values for that variables, but of course only when the corresponding xi is not missing itself.

(Refer Slide Time: 03:20)

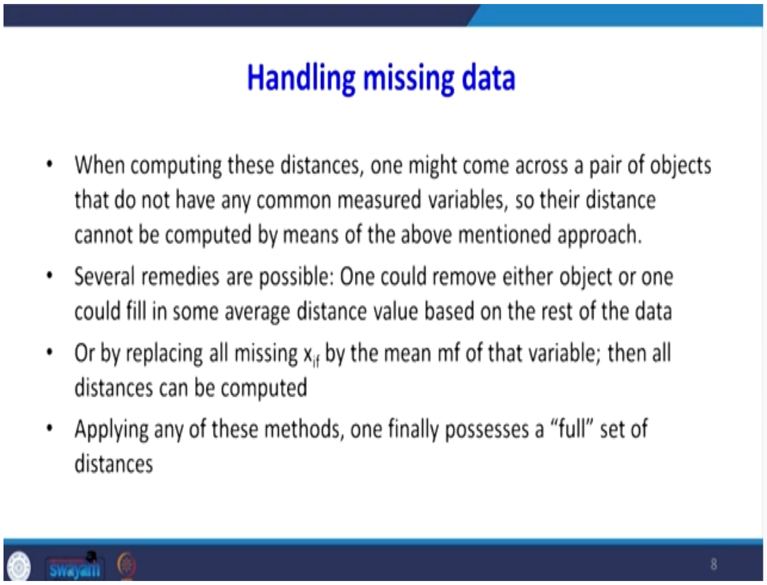


Handling missing data

- In the computation of distances (based on either the x_i, or the z_i,) similar precautions must be taken
- When calculating the distances d(i, j), only those variables are considered in the sum for which the measurements for both objects are present subsequently the sum is multiplied by p and divided by the actual number of terms (in the case of Euclidean distances this is done before taking the square root)
- Such a procedure only makes sense when the variables are thought of as having the same weight (for instance, this can be done after standardization)

In the computation of distances based on the either Xi or the Zi similar precautions must be taken when calculating the distances d of i, j only those variables are considered in the sum of which the measurements of both objects are present. Subsequently, the sum is multiplied by p and divided by the actual number of terms. In the case of Euclidean distances, this is done before taking the square root. Such a procedure only make sense when the variables are thought of as having the same weight. For instance, this can be done after standardization.

(Refer Slide Time: 04:03)

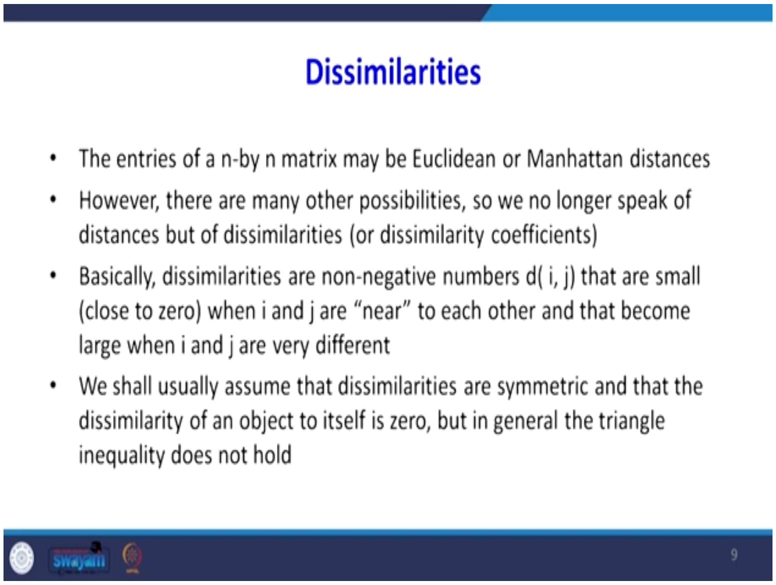


Handling missing data

- When computing these distances, one might come across a pair of objects that do not have any common measured variables, so their distance cannot be computed by means of the above mentioned approach.
- Several remedies are possible: One could remove either object or one could fill in some average distance value based on the rest of the data
- Or by replacing all missing x_{if} by the mean mf of that variable; then all distances can be computed
- Applying any of these methods, one finally possesses a “full” set of distances

When computing these distances, one might come across a pair of objects that do not have any common measured variables, so their distance cannot be computed by means of above-mentioned approach. Several remedies are possible: One could remove either object, or one could feel some average distance value based on the rest of the data. Or, by replacing all missing xif by the mean of mf, that variable, then all distances can be computed. Applying any of these methods one finally possesses a full set of distances.

(Refer Slide Time: 04:44)

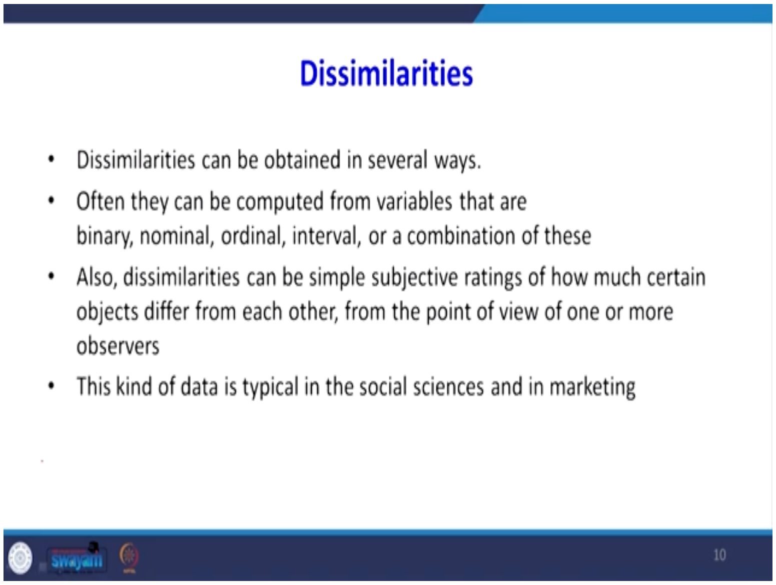


Dissimilarities

- The entries of a n-by n matrix may be Euclidean or Manhattan distances
- However, there are many other possibilities, so we no longer speak of distances but of dissimilarities (or dissimilarity coefficients)
- Basically, dissimilarities are non-negative numbers d(i, j) that are small (close to zero) when i and j are "near" to each other and that become large when i and j are very different
- We shall usually assume that dissimilarities are symmetric and that the dissimilarity of an object to itself is zero, but in general the triangle inequality does not hold

Then we will go to another topic that is dissimilarities. The entries of n-by-n matrix may be Euclidean or Manhattan distances. However, there are many other possibilities, so we no longer speak of distances, but dissimilarities or dissimilarity coefficients. Basically, dissimilarities are non- negative numbers that is d of i, j that are small, close to 0 when i and j are near to each other and they become large when i and j are very different. We shall usually assume that dissimilarities are symmetric and that the dissimilarity of an object to itself 0. But in general, the triangle inequality does not hold.

(Refer Slide Time: 05:40)



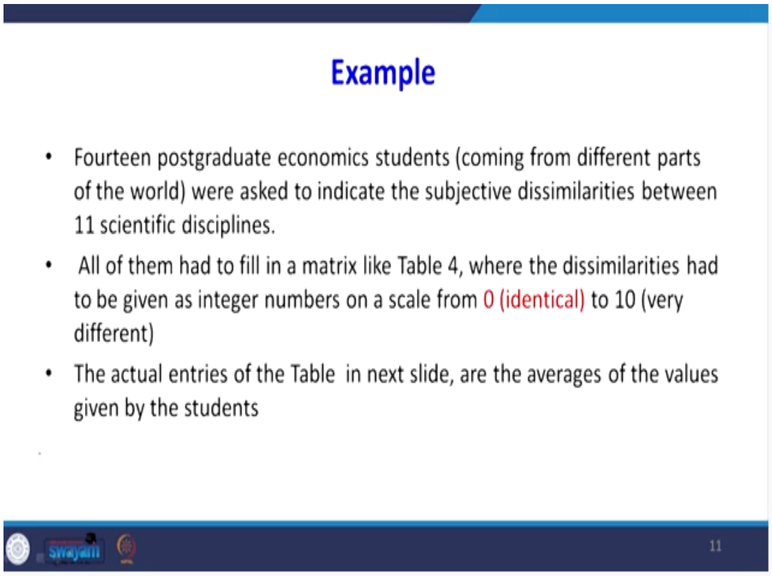
Dissimilarities

- Dissimilarities can be obtained in several ways.
- Often they can be computed from variables that are binary, nominal, ordinal, interval, or a combination of these
- Also, dissimilarities can be simple subjective ratings of how much certain objects differ from each other, from the point of view of one or more observers
- This kind of data is typical in the social sciences and in marketing

10

Dissimilarities can be obtained in several ways. Often, they can be computed from variables that are binary, nominal, ordinal, interval or combination of these. Also dissimilarities can be simple subjective rating of how much certain objects differ from each other from the point of view of one or more observers. This kind of data is typical in the social science or in the marketing.

(Refer Slide Time: 06:11)



Example

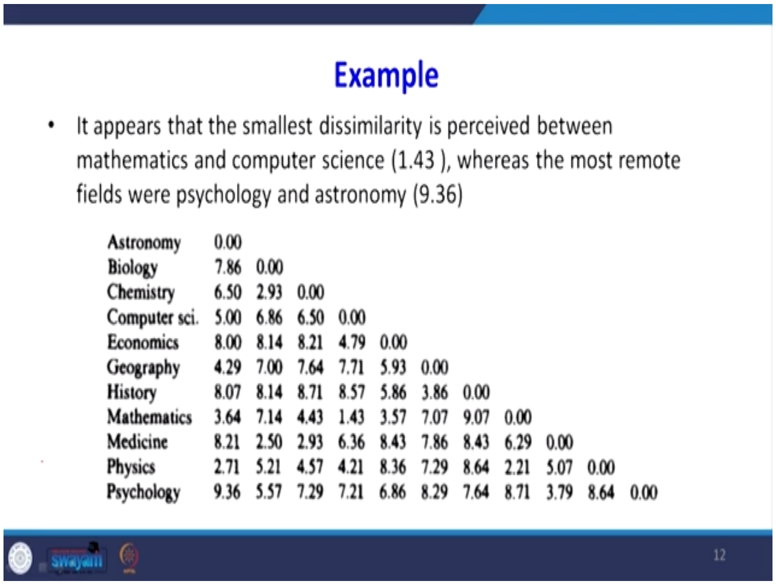
- Fourteen postgraduate economics students (coming from different parts of the world) were asked to indicate the subjective dissimilarities between 11 scientific disciplines.
- All of them had to fill in a matrix like Table 4, where the dissimilarities had to be given as integer numbers on a scale from 0 (identical) to 10 (very different)
- The actual entries of the Table in next slide, are the averages of the values given by the students

11

Let us take an example then I will explain the concept of dissimilarities fourteen post-graduate economic students coming from different parts of the world were asked to indicate the subjective dissimilarities between 11 scientific disciplines. All of them had to fill in a matrix, like in table 4 in the next slide where the dissimilarities had to be given as integer numbers, on a scale of 0 to

10, where the 0 represents identical 10 represents very different. The actual entries of the table in the next slides are the average of these values given by the students.

(Refer Slide Time: 06:54)



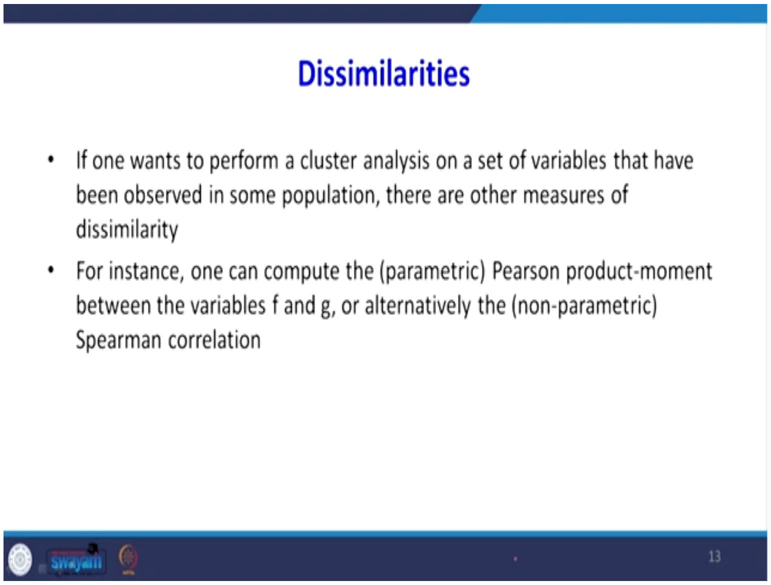
Example

- It appears that the smallest dissimilarity is perceived between mathematics and computer science (1.43), whereas the most remote fields were psychology and astronomy (9.36)

Astronomy	0.00										
Biology	7.86	0.00									
Chemistry	6.50	2.93	0.00								
Computer sci.	5.00	6.86	6.50	0.00							
Economics	8.00	8.14	8.21	4.79	0.00						
Geography	4.29	7.00	7.64	7.71	5.93	0.00					
History	8.07	8.14	8.71	8.57	5.86	3.86	0.00				
Mathematics	3.64	7.14	4.43	1.43	3.57	7.07	9.07	0.00			
Medicine	8.21	2.50	2.93	6.36	8.43	7.86	8.43	6.29	0.00		
Physics	2.71	5.21	4.57	4.21	8.36	7.29	8.64	2.21	5.07	0.00	
Psychology	9.36	5.57	7.29	7.21	6.86	8.29	7.64	8.71	3.79	8.64	0.00

It appears that the smallest dissimilarity is perceived between mathematics and computer science that value is 1.43 mathematics and computer science. This is our smallest dissimilarity whereas the most remote fields where psychology and astronomy psychology, astronomy. So this table represents dissimilarity matrix from that we can directly read, which is having lesser dissimilarity, which is having more dissimilarity.

(Refer Slide Time: 07:35)

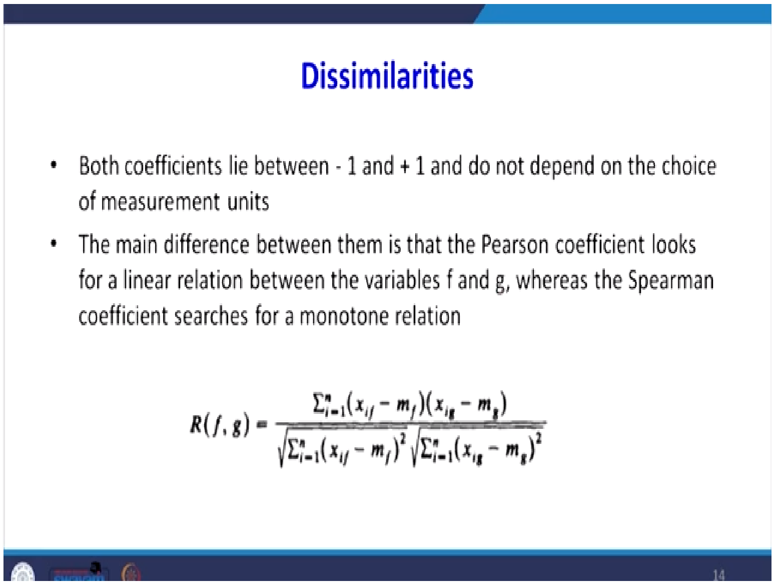


Dissimilarities

- If one wants to perform a cluster analysis on a set of variables that have been observed in some population, there are other measures of dissimilarity
- For instance, one can compute the (parametric) Pearson product-moment between the variables f and g, or alternatively the (non-parametric) Spearman correlation

If one wants to perform a cluster analysis on a set of variables that have been observed in some population. There are other measures of dissimilarity. For instance, one can compute the Parametric Pearson product-moment between the variables f and g or alternatively non-parametric spearman correlation. Here the dissimilarity can be found with the help of your Pearson correlation or spearmen correlation. We know that the Pearson correlation is a parametric method spearmen correlation is non-parametric method. Because Spearman correlation is applicable only for ordinal data.

(Refer Slide Time: 08:18)



Dissimilarities

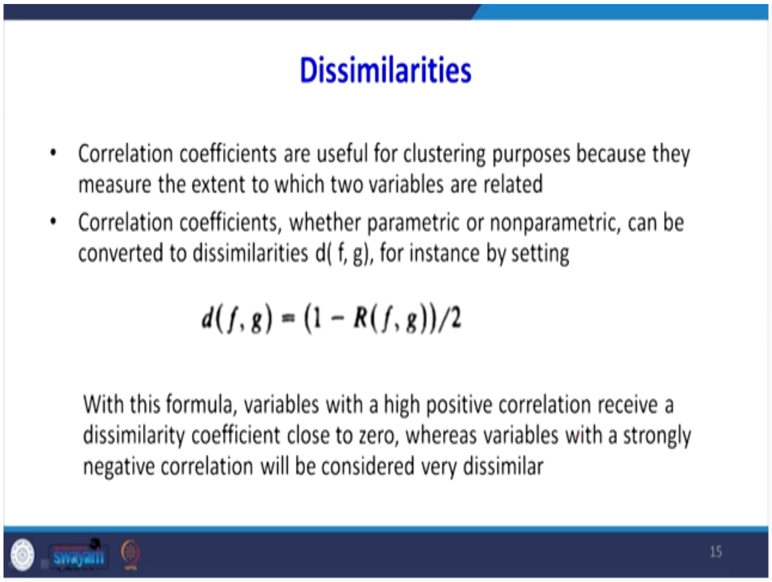
- Both coefficients lie between - 1 and + 1 and do not depend on the choice of measurement units
- The main difference between them is that the Pearson coefficient looks for a linear relation between the variables f and g, whereas the Spearman coefficient searches for a monotone relation

$$R(f,g)=\frac{\sum_{i=1}^{n}(x_{if}-m_f)(x_{ig}-m_g)}{\sqrt{\sum_{i=1}^{n}(x_{if}-m_f)^2}\sqrt{\sum_{i=1}^{n}(x_{ig}-m_g)^2}}$$

Both the coefficients lay between - 1 and + 1. Which one I am saying where our Pearson and Spearman correlation and do not depend on the choice of measurement units. We need not bother about the units because we are going to see the range of correlation coefficient is – 1 to + 1. Similarly, the Spearman correlation value also between -1 to + 1. That value does not depending upon what type of units of the data.

The main difference between is that the Pearson coefficients look for a linear relationship between variables f and g, whereas the spearmen coefficient searches for monotone relations. So, this is formula for our correlation coefficient, so we call it as r the correlation coefficient we have studied this formula already. So the correlation coefficient row is this is nothing but co variance, co variance of x, y divided by standard deviation of x and standard deviation of y. So this is in some other format this is x, y the correlation coefficient, not x, y here you can call it as f, g.

(Refer Slide Time: 09:40)



Dissimilarities

- Correlation coefficients are useful for clustering purposes because they measure the extent to which two variables are related
- Correlation coefficients, whether parametric or nonparametric, can be converted to dissimilarities d(f, g), for instance by setting

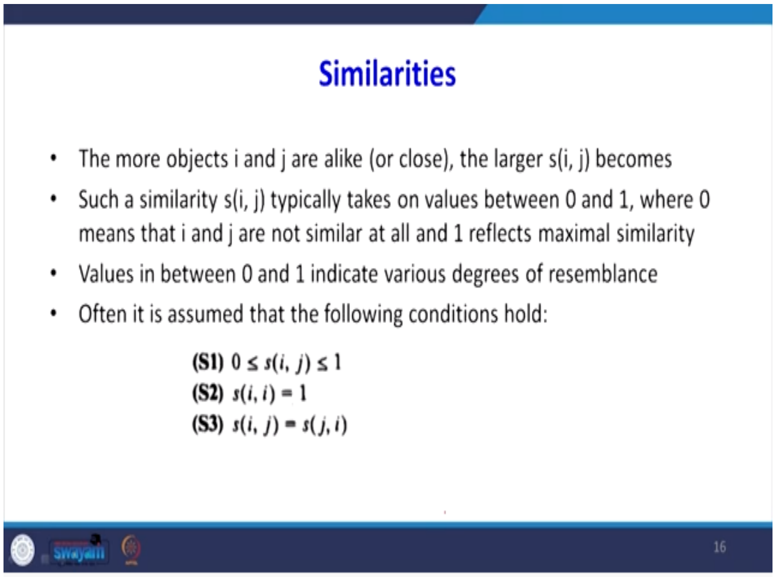
$$d(f, g) = (1 - R(f, g))/2$$

With this formula, variables with a high positive correlation receive a dissimilarity coefficient close to zero, whereas variables with a strongly negative correlation will be considered very dissimilar

Correlation coefficients are useful for clustering purposes because they measures the extent to which two variables are related. Correlation coefficients, whether parametric or non-parametric can be converted into dissimilarities d(f, g) for instance, by setting by this relationship. So dissimilarity between object (f, g)= (1- R that is correlation coefficient between (f, g)) divided by 2. ith this formula variables with a high positive correlation receive a dissimilarity coefficient close to zero whereas the variables with a strongly negative correlation will be considered as very dissimilar.

Why this kind of conversion is required the range of dissimilarity is between 0 to 1, but sometime what will happen the value of correlation coefficient between – 1 to + 1. So convert into to 0 to 1 scale we can use this transformation.

(Refer Slide Time: 10:46)



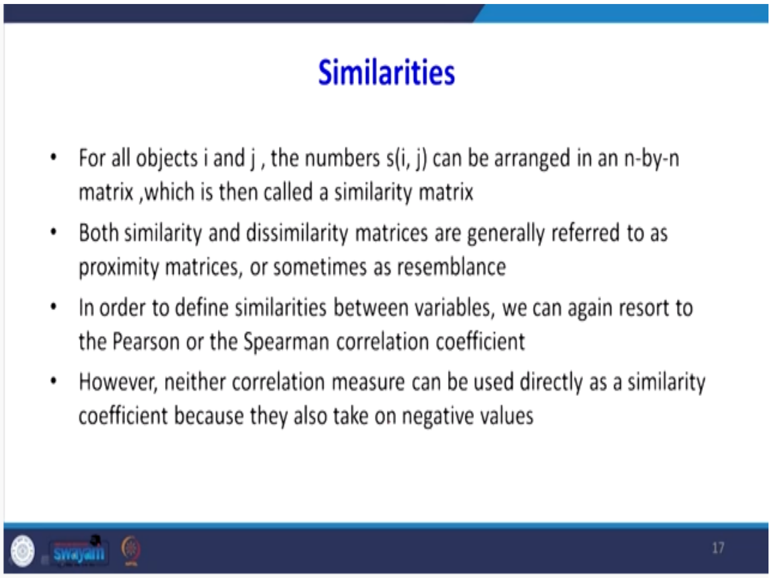
Similarities

- The more objects i and j are alike (or close), the larger s(i, j) becomes
- Such a similarity s(i, j) typically takes on values between 0 and 1, where 0 means that i and j are not similar at all and 1 reflects maximal similarity
- Values in between 0 and 1 indicate various degrees of resemblance
- Often it is assumed that the following conditions hold:

(S1) $0 \leq s(i, j) \leq 1$
(S2) $s(i, i) = 1$
(S3) $s(i, j) = s(j, i)$

Now we enter into another concept called similarities. Previously we are explaining about dissimilarity and how we are going to study about what is similarities. The more objects and j are alike, so the larger will be similarity between s (i, j) becomes. Such a similarity s of (i, j) typically takes on values between 0 to 1 whereas 0 means that i and j are not similar at all, and 1 reflects maximum similarity. Values between 0 and 1 indicate various degrees of resemblance. Often it is assumed that the following conditions hold. So, S1: 0 <= s (i, j) <= 1, because the range of similarities between 0 to 1. S2: the similarity between i, i itself 1, the similarity s (i, j) = s(j, i)

(Refer Slide Time: 11:48)



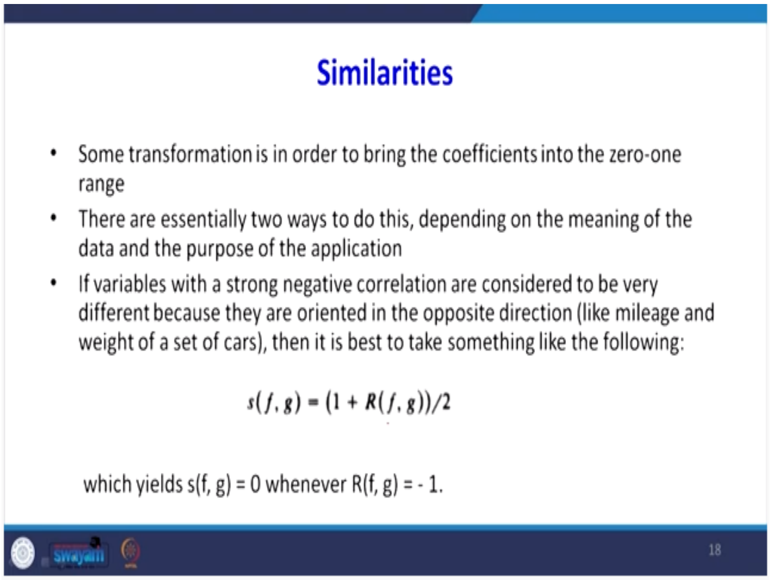
Similarities

- For all objects i and j , the numbers s(i, j) can be arranged in an n-by-n matrix ,which is then called a similarity matrix
- Both similarity and dissimilarity matrices are generally referred to as proximity matrices, or sometimes as resemblance
- In order to define similarities between variables, we can again resort to the Pearson or the Spearman correlation coefficient
- However, neither correlation measure can be used directly as a similarity coefficient because they also take on negative values

We will continue the concept of similarities for all objects i and j the numbers s of (i, j) can be arranged in an n-by-n matrix which is then called similarity matrix. Both similarity and dissimilarity matrices are generally referred to as proximity matrices sometimes as a resemblance. In order to define similarities between variables, we can again resort to a Pearson or Spearman correlation coefficient.

However, neither correlation measures can be used directly as a similarity coefficient because they also take on negative values because we cannot take the value of correlation and Spearman correlation as it is because they may range between - 1 to + 1, but the similarity values between 0 to 1.

(Refer Slide Time: 12:41)



Similarities

- Some transformation is in order to bring the coefficients into the zero-one range
- There are essentially two ways to do this, depending on the meaning of the data and the purpose of the application
- If variables with a strong negative correlation are considered to be very different because they are oriented in the opposite direction (like mileage and weight of a set of cars), then it is best to take something like the following:

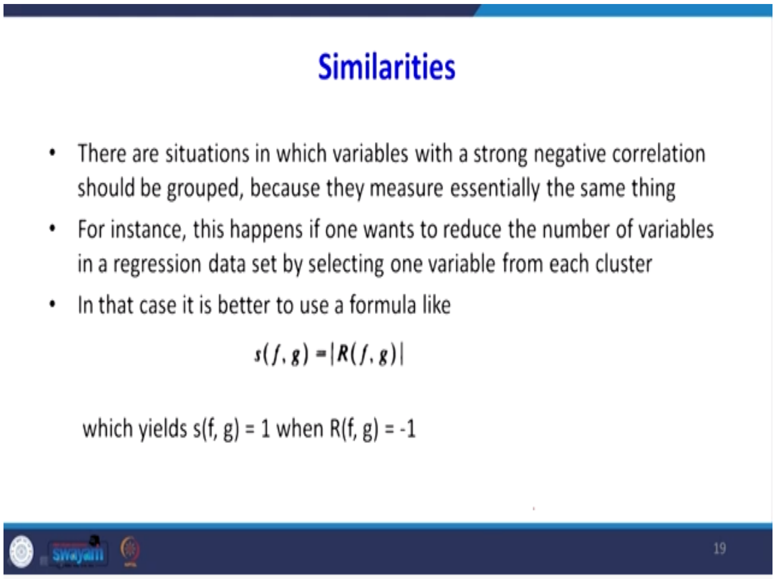
$$s(f, g) = (1 + R(f, g))/2$$

which yields s(f, g) = 0 whenever R(f, g) = - 1.

So in that case, we have to go for some transformation, some transformation is in order to bring the coefficients into the zero-one range. There are essentially two ways to do this, depending on the meaning of the data and the purpose of the application. If the variables with a strong negative correlation are considered to be very different because they are oriented in the opposite direction, like mileage and weight of a set of cars, then it is best to take something like the following.

You have to follow this transformation s of (f, g) = (1 + R of (f, g))/2 what will happen here? We have added some constant so that constant will nullify the negative effect which yields the similarity between f and g = 0 whenever the correlation coefficient is - 1 because - 1 and + 1 becomes 0. So this take care that the similarity value comes between 0 to 1.

(Refer Slide Time: 13:49)



Similarities

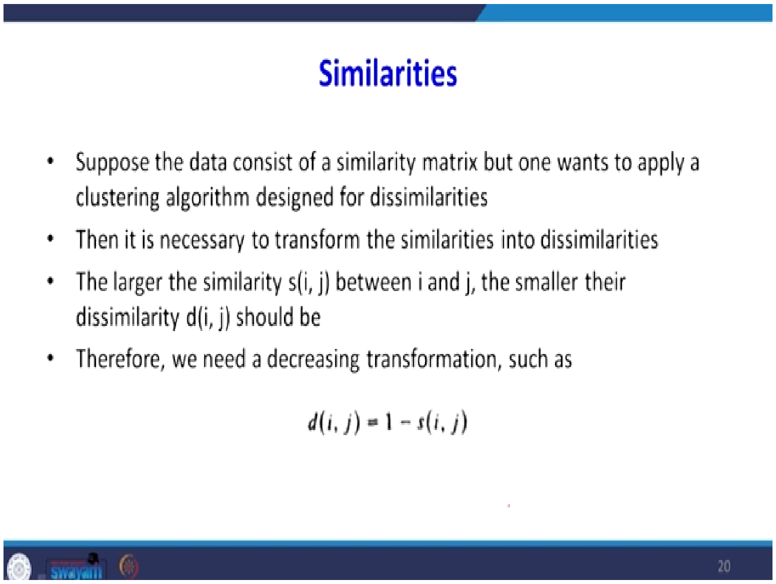
- There are situations in which variables with a strong negative correlation should be grouped, because they measure essentially the same thing
- For instance, this happens if one wants to reduce the number of variables in a regression data set by selecting one variable from each cluster
- In that case it is better to use a formula like

$$s(f,g) = |R(f,g)|$$

which yields s(f, g) = 1 when R(f, g) = -1

There are situations in which variables with a strong negative correlation should be grouped because they measure essentially the same thing. For instance, this happens if one wants to reduce the number of variables in a regression dataset by selecting one variable from each cluster. In that case, it is better to use formula like this. Similarity between (f, g) = the modulus value of correlation coefficient between f and g, which yields that the similarity between (f, g) = 1 when the correlation coefficient is -1. We have to take only the positive values.

(Refer Slide Time: 14:31)



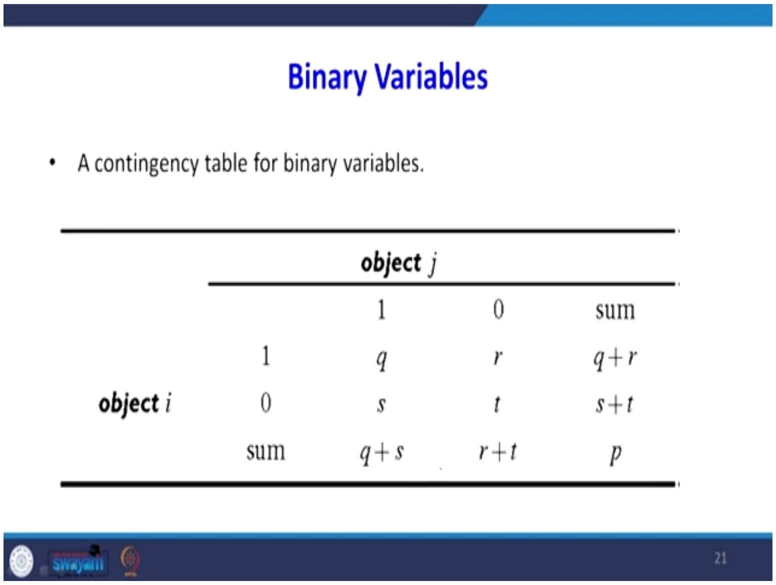
Similarities

- Suppose the data consist of a similarity matrix but one wants to apply a clustering algorithm designed for dissimilarities
- Then it is necessary to transform the similarities into dissimilarities
- The larger the similarity s(i, j) between i and j, the smaller their dissimilarity d(i, j) should be
- Therefore, we need a decreasing transformation, such as

$$d(i,j) = 1 - s(i,j)$$

Suppose the data consist of similarity matrix, but one wants to apply a clustering algorithm designed for dissimilarities. Then it is necessary to transform the similarities into dissimilarities. The larger the similarity the similarity between s (i, j), between i and j the smaller their dissimilarity d of (i, j) should be. Therefore, we need a decreasing transformation. This is a very important result. So what it says that if you want to know the dissimilarity between two objects i, j that is nothing but 1 - similarity between i, j

(Refer Slide Time: 15:13)



Binary Variables

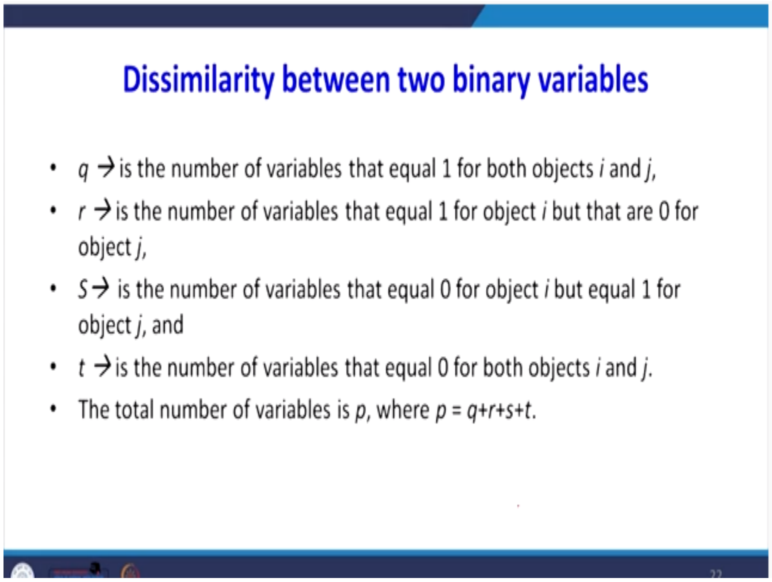
- A contingency table for binary variables.

		object *j*		
		1	0	sum
object *i*	1	q	r	$q+r$
	0	s	t	$s+t$
	sum	$q+s$	$r+t$	p

Let us take a binary type variable for that let us find the similarity and dissimilarity. Suppose a contingency table for binary variable is given. There is an object i and object j you see that object i is 1 0 there are two possibility object j also 1 0. So the q represents where the object i also takes value 1 object j also takes value 1. This r represents, q is the number of values here r represents i = 1 j = 0 this s represents number of values where i = 0 j = 1, t represents both i and j = 0.

The row sum is q + r for when i =1, when i = 0 the row sum is s +t same thing the column sum ‘s’, when j = 1 the column sum is q + s, when j = 0 the column sum is r + t. So the sum of q, r, s, t that is nothing but your value p.

(Refer Slide Time: 16:30)

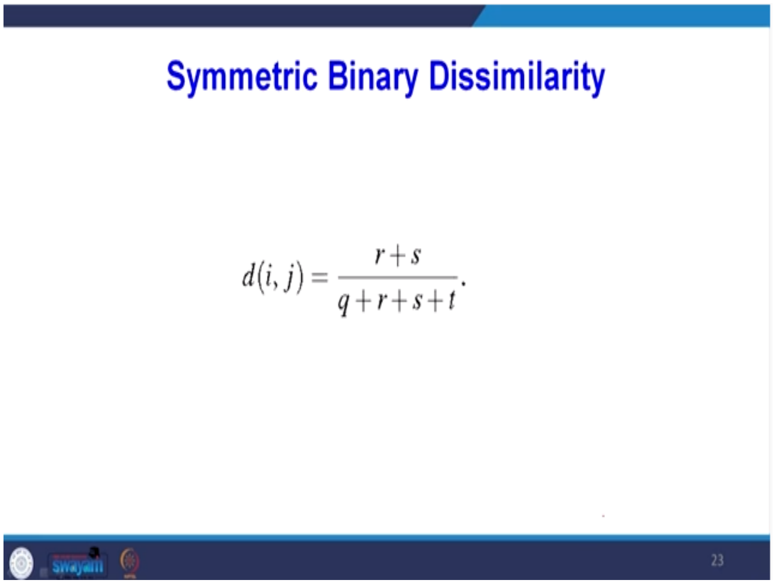


Dissimilarity between two binary variables

- *q* → is the number of variables that equal 1 for both objects *i* and *j*,
- *r* → is the number of variables that equal 1 for object *i* but that are 0 for object *j*,
- *S* → is the number of variables that equal 0 for object *i* but equal 1 for object *j*, and
- *t* → is the number of variables that equal 0 for both objects *i* and *j*.
- The total number of variables is *p*, where *p = q+r+s+t*.

What is the meaning of this q, r, s, t? q represents the number of variables that equa1 1 for both objects i and j you see that q is the number of variables r is the number of variables that equal one for object i but that are 0 for object j. S represents number of variables that equals 0 for object i, but equal 1 for object j. So t represents the number of variables that equals 0 for both objects i and j. The total number of variables is p where p = q + r + s + t.

(Refer Slide Time: 17:11)

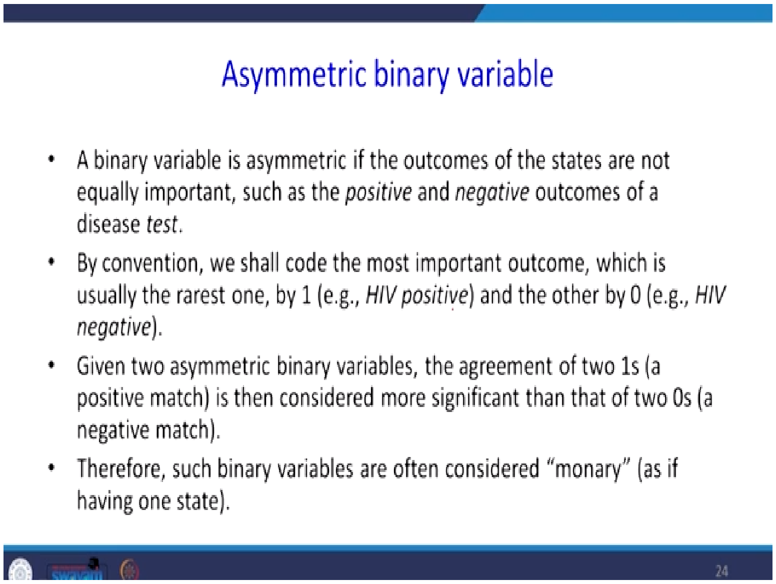


Symmetric Binary Dissimilarity

$$d(i,j) = \frac{r+s}{q+r+s+t}.$$

The dissimilarity between symmetric binary variable, so from the previous table. What is the meaning of Symmetric binary variable is example is gender suppose 0 Male 1 female. You can reverse the code also there would not be any problem on this. So that is example of Symmetric Binary variable. So for Symmetric binary variable, how to find out dissimilarity. So dissimilarity between i, j is it is r +s what is r+ s we will go this. So this one, the dissimilarity is this value r + s divided by sum of the all values r + s divided by q + r + s + t.

(Refer Slide Time: 18:02)



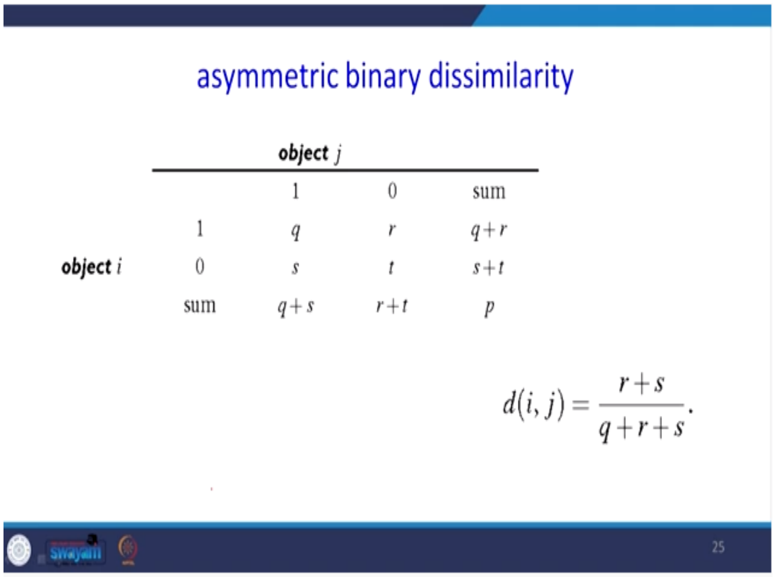
Asymmetric binary variable

- A binary variable is asymmetric if the outcomes of the states are not equally important, such as the *positive* and *negative* outcomes of a disease *test*.
- By convention, we shall code the most important outcome, which is usually the rarest one, by 1 (e.g., *HIV positive*) and the other by 0 (e.g., *HIV negative*).
- Given two asymmetric binary variables, the agreement of two 1s (a positive match) is then considered more significant than that of two 0s (a negative match).
- Therefore, such binary variables are often considered “monary” (as if having one state).

Then let us see what is the meaning of Asymmetric binary variable. A binary variable is Asymmetric. If the outcomes of the states are not equally important, such as positive and negative outcome of disease test. By convention, we shall code the most important outcome, which is usually the rarest one by 1. For example, HIV positive that is the rarest one we will code it as 1 and the other by 0 HIV negative.

Given two Asymmetric binary variable the agreement of two 1s that is a positive match is considered more significant than that of two 0s that is negative match. Therefore, such binary variable are often considered as monary as if having only one state because we need not bother about the zero state, because zero state is that non-presence of HIV, because we are more concerned about presence of HIV where the state of one is more important.

(Refer Slide Time: 19:07)



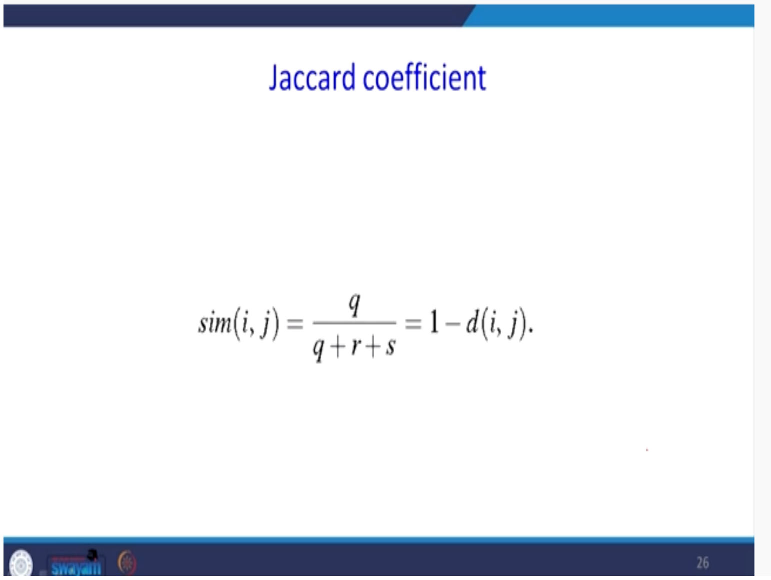
asymmetric binary dissimilarity

		object j		
		1	0	sum
	1	q	r	$q+r$
object i	0	s	t	$s+t$
	sum	$q+s$	$r+t$	p

$$d(i,j)=\frac{r+s}{q+r+s}.$$

Now let us see how to find out dissimilarity value between Asymmetric binary variable. The same table which I have given contingency table which I have given previous table I have given. So the dissimilarity between Asymmetric matrixes d of (i, j). So we are considered about only r and s in this in the denominator there would not be t because we are not considering 0. So only we are writing q+ r + s. If it is Symmetric binary dissimilarity, the difference is there was a t element was there here. But here in the asymmetric binary dissimilarity formula there is no ‘t’ element.

(Refer Slide Time: 19:51)

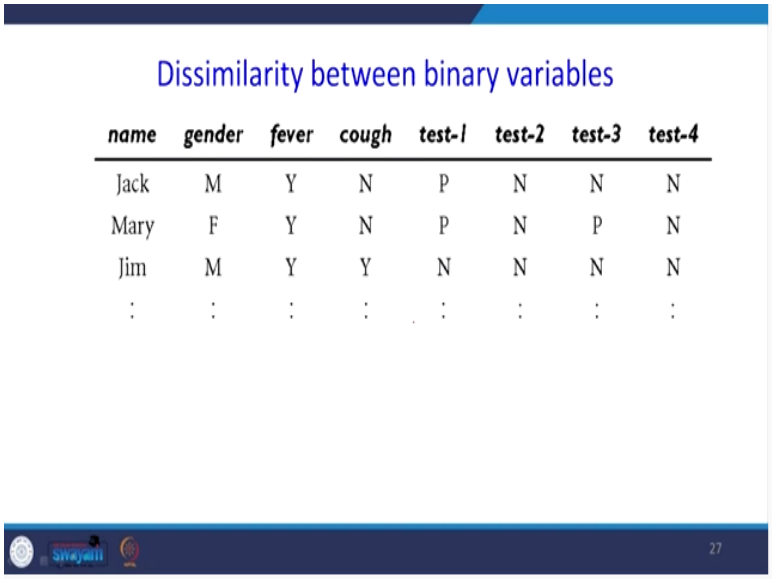


Jaccard coefficient

$$sim(i,j)=\frac{q}{q+r+s}=1-d(i,j).$$

Even we have seen this relationship, that relationship called the Jaccard co-efficient. That is the similarity between(i, j) = 1 - dissimilarity. So similarity q divided by (q + r+ s), where is q this one, we are bothered about only the pretense of 1 so q divided by q + r + s that is your similarity between i, j for a asymmetric binary variable. So if we want to know dissimilarity, that is simply the similarity equal to 1 minus dissimilarity between i and j.

(Refer Slide Time: 20:32)

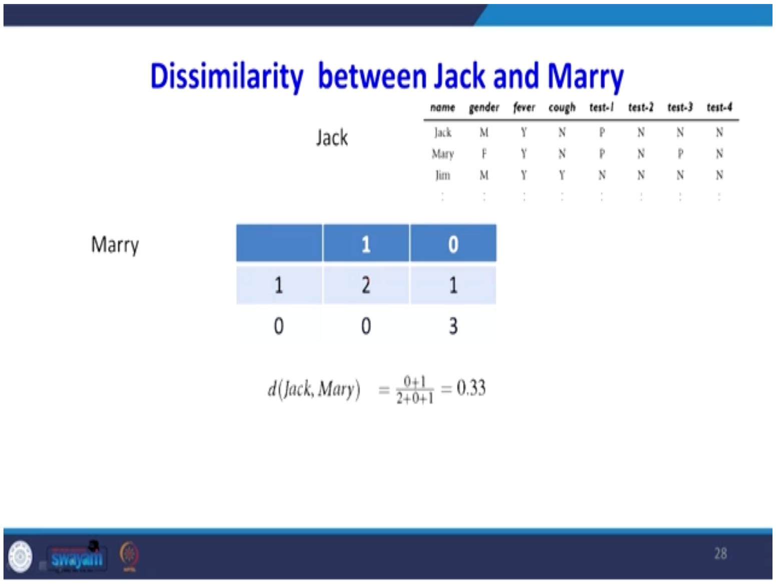


Dissimilarity between binary variables

name	gender	fever	cough	test-1	test-2	test-3	test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	Y	N	N	N	N
:	:	:	:	:	:	:	:

Now let us take an example we will find out for Asymmetric binary variable how to find out dissimilarity matrix. This table shows there are different name is there Jack, Mary, Jim. There is a gender here gender is Symmetric binary variable. We are not going to consider this one because this is a different test fever, cough test 1, test 2, test 3, test 4. This is Asymmetric variable because where the presence of 1 is more important Y represents and P represents 1, N represents 0.

(Refer Slide Time: 21:19)



Dissimilarity between Jack and Marry

name	gender	fever	cough	test-1	test-2	test-3	test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	Y	N	N	N	N
:	:	:	:	:	:	:	:

Jack

Marry

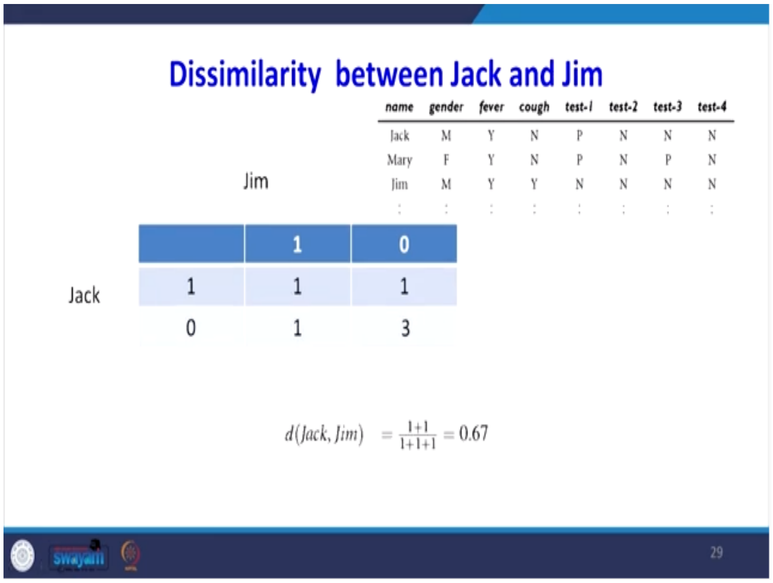
	1	0
1	2	1
0	0	3

$$d(Jack, Mary) = \frac{0+1}{2+0+1} = 0.33$$

For this matrix, let us find the dissimilarity matrix between Jack and Mary. So I brought the table again we will let us find out the dissimilarity matrices between Jack and Mary. So for Mary, there are two possibility 1 0 for Jack that is under 2 possibility, 1 and 0. So let us count how we got this 2 so Mary also 1 Jack also 1 there are two possibilities there Mary this is 1 possibility this is another possibility. So there is a two count, so we have written it as 2.

Then how we got this 1? Mary is 1 jack is 0 so that means this one where Mary is 1 Jack is 0. Now we will go this column where Mary is 0 Jack is 1 so Mary 0 is this one, I think there is no value for this. Let us see the last option that is Mary also 0 Jack is also 0, So this 1 2 this is 3 that is 3. So if you want to know the dissimilarity distance between Jack and Mary, so we know that the formula is so we will add this 0 + 1 divided by 2 + 0 + 1 so we got 0.33.

(Refer Slide Time: 22:50)



Dissimilarity between Jack and Jim

name	*gender*	*fever*	*cough*	*test-1*	*test-2*	*test-3*	*test-4*
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	Y	N	N	N	N
:	:	:	:	:	:	:	:

Jim

Jack

	1	0
1	1	1
0	1	3

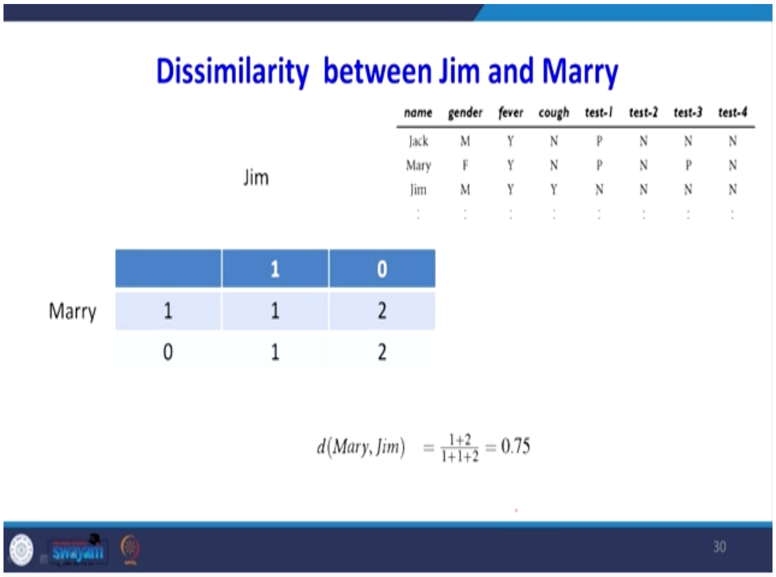
$$d(Jack, Jim) = \frac{1+1}{1+1+1} = 0.67$$

29

Similarly, now let us find how to find out the dissimilarity between Jack and Jim. So Jack is taken in rows Jim is taken as in the column. So first we will find out how we got this 1. So this case is Jack also 1 Jim also 1. So this category so Jack also 1 Jim also 1. I think there is only one possibility, so it is 1 how we got this 1, the Jack is 1 Jim is 0. So this value Jack is 1 presence Jim is 0 that is 1. So how we got this1 where Jack is 0 Jim is 1. So Jack is 0 yeah, this value Jack is 0 no means 0, Y means 1.

Let us see how we got this value 3 so Jack also is 0 Jim also is 0 so this 1 no this one 1, 2, 3 that is how we got the 3 values. So if you want to know the dissimilarity between Jack and Jim. So this is 1 + 1 divided by 1 + 1 + 1 so 2/3 it is 0.67.

(Refer Slide Time: 24:16)



Dissimilarity between Jim and Marry

name	gender	fever	cough	test-1	test-2	test-3	test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	Y	N	N	N	N
:	:	:	:	:	:	:	:

		Jim	
		1	0
Marry	1	1	2
	0	1	2

$$d(Mary, Jim) = \frac{1+2}{1+1+2} = 0.75$$

Let us take another example the dissimilarity between Jim and Mary. So Mary there is two option 1 and 0. But Jim also there are two options 1 and 0 let us see how we got this value 1. Mary is 1, Jim also 1. So this possibility this one the second case Mary is 1, Jim is 0 this is one value. Mary is 1 Jim is 0, so there are two possibility. So that is where we got the value 2 how we got to this value 1, Mary is 0 Jim is 1, so Mary is 0 Jim = 1 that is this value.

So how we got this 2 Mary is also 0 Jim also 0 so these two possibilities, Mary also 0 Jim also 0 test 1 here also Mary is 0 Jim also 0. So if you want to know asymmetric dissimilarity between Jim and Mary it is 1 + 2 this value plus this one divided by this 1 +1 + 2 that is 4 so we got 0.75. So this is the way to find out asymmetric dissimilarity between different variables. In this class, we have seen how to handle the missing data for cluster analysis.

Then I have explained the concept of similarity and dissimilarity matrix. Then we have studied symmetric and asymmetric binary variable and how to find out the dissimilarity between symmetric binary variables and dissimilarity between asymmetric binary variables. Thank you.