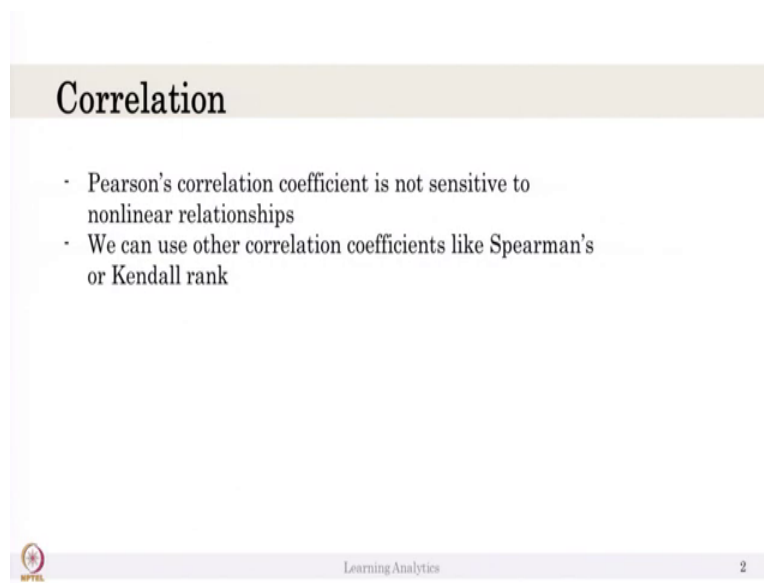


Learning Analytics Tools
Professor Ramkumar Rajendran
Educational Technology
Indian Institute of Technology, Bombay
Lecture 5.4
Spearman's Rank Correlation


In this video, let us look at what is Spearman's Rank Correlation because we know the limitations of Pearson correlation coefficient.

(Refer Slide Time: 00:26)



Correlation

- Pearson's correlation coefficient is not sensitive to nonlinear relationships
- We can use other correlation coefficients like Spearman's or Kendall rank


 Learning Analytics 2

So, we see that Pearson's correlation coefficient is not sensitive to nonlinear relationships also for this kind of data, we can use other correlation coefficients like Spearman's or Kendall rank. Let us look at Spearman's in this course. I am not going to discuss Kendall's rank because there will be too much of talking about correlation only in this course.


(Refer Slide Time: 00:53)

Spearman's rank correlation

- Pearson correlation coefficient between ranks
- Values between -1 to +1
 - - (negative) sign indicates indirect relationship of X and Y
 - + (positive) direct relations $x \uparrow \rightarrow y \uparrow$
 - Values indicate the magnitude of relation between X and Y



$x \uparrow \rightarrow y \downarrow$



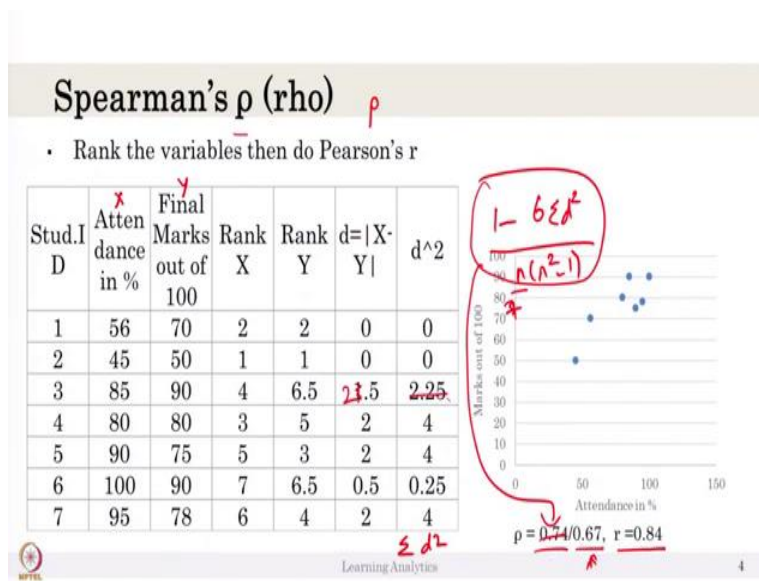
Learning Analytics

3

But let us have a feel of what is Spearman's rank correlation. Similar to Pearson's correlation coefficient, Spearman's rank Correlation also varies between plus 1 to minus 1 and exactly it is a negative sign indicates there is an indirect relationship between X and Y, if X increases Y will decrease and if X increases then Y will decrease.

And the positive sign indicates if X increases, then Y also have an increase. So, directly proportional and the values indicates the magnitude of relationship between X and Y. If value is more like 0.8 or 0.9 is strongly correlated, 0.2 indicates the correlation is bit weak. The first point is, Pearson's correlation coefficient between ranks is called Spearman's rank correlation. Spearman's rank correlation is nothing but Pearson's correlation coefficient applied on ranks on the variables we use. We will discuss that in detail, in the coming slides.

(Refer Slide Time: 01:58)



So, let us look at what spearman's rho is. Let us look at seven students data and we have attendance and we are find marks out of this. Now first thing we have to do is let us take the, this is X and this is the Y. We will do the ranking of the available data. Like this is the least and maybe arrange it in ascending order and this is the first rank, this is second rank, first is 45, second is 56 then 80, then 85, 90, 95 100 those are ranks.

Similarly, for the marks, we can rank this is the least value very high rank. So, 1 and next is 2 and this is 3 and 4, 5, there are two 90, so we do not give 6 and 7, we will take the average of 6 and 7, and put the value 6.5. Since these variables are ranked. If you compute Pearson's, r on this this, that gives you the Spearman's rho.

There is other's formula instead of using Pearson's on this rank, you compute the difference between these two ranks, so the difference between these two rank 0, this is 1.5 etc

So, the difference is 2.5 and it is 2 and it is 2 and this is 0.5 and this is 2. So the values will be if you I just did the mod because I just if it is negative also consider it is positive because I just want absolute difference not the sign. You will calculate the difference for all values.

We want to consider the square of this value. So, whatever the difference either negative or positive does not matter. So, we get 0, 0, 0, 2.25, 4, 4, 0.25 and 4. So, you can use these values the summation of these d square values and there is a formula to compute a Spearman's rho it is

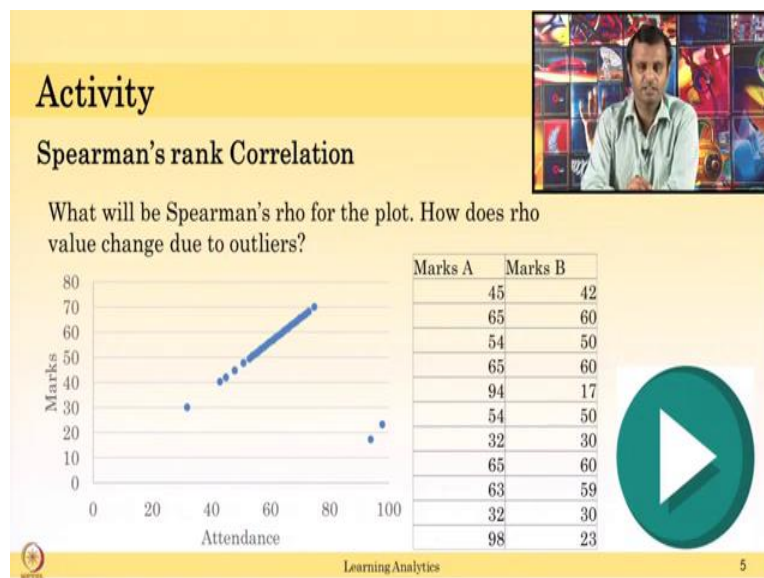
actually the 1 minus 6 into summation of d square values n into n square minus 1, n is actually number of samples .

And we can use this formula to compute Spearman's rho and this formula is a simplified formula of doing Pearson's or on the rank. So this is how you compute Spearman's rho. So, let us look at the plot the plot looks like this for this variables and it tells tells a linear relationship. There is a good relationship from here to here.

Let us look at the Spearman's rho value and our value. The rho is 0.74, this not correct because I used 1.5 instead of 2.5 if you compute it you get different value.

But this is how the Spearman's rho is computed that is Spearman's rho is computed by using rank the variables then apply Pearson's r on the rank. So, here r indicates Pearson's correlation coefficient 0.84 it is a high correlation, but Spearman's is not saying it is high correlated because the relationship may not be linear or something is missing. So, let us see what is the exact difference between rho and r in a next slide.

(Refer Slide Time: 07:45)



In this activity I just want you to show how Spearman's rho is sensitive to the outliers. So, now we saw this particular values in a previous slides, I changed two values that is for marks A, I change in this to 17 and for the high marks in A, I changed this 23, so I made it two outliers out of this mark I just change myself to create two outliers because every other value is perfectly

high score means high scoring B high scoring A a but I just change these two values to show the outlier.

Can you predict or guess, you do not need to compute it and mathematically, what will be this Spearman's rho for this plot? Also what will be the r value Pearson's correlation coefficient this plot, and how it will differ? Can you guess it? And how it will vary for the outliers in everything? If you have done that please resume the video to continue.


(Refer Slide Time: 08:48)


Activity

Spearman's rho

- $r = 0.34$
- $\rho = 0.81$
- The rank value is used to compute the correlation

$$\begin{array}{r|l} 94 - 59 & 17 - 1 = 58^2 \\ 98 - 60 & 23 - 2 = 58^2 \end{array}$$
$$\frac{60(60-1)}{2}$$



 Learning Analytics 6

So the Spearman's rho for that particular is 0.81, but correlation coefficient r is 0.34, because it is due to outlier the line tries to fit all the lines and there is a huge difference between the points and the line. So the r value is low, but rho is good, rho is not sensitive to that outlier. The rank value is used to compute the correlation why Spearman's rho is not much sensitive. It is because out of all 60 students data only 2 data is not correct. For example, in that example 94 and 98. So this will be highly ranked 59 and 60 and it is 17 and 23.

So, when you compute the difference between ranks, it might be say 58 and 58 and square of these values might be huge and but when compared to the other values there are other 58 values, which is closely related. I leave like all the ranks are might be equal except these ranks and so one or two value, So there is value and if you divide by 60 into 60 square minus 1, this value will be negligible. So, that Spearman's rho is not much sensitive to the outliers like the Pearson's correlation coefficient.

So we try to compute the correlation coefficient as much as good based on how many numbers or how many values are highly correlated and it might ignore the outliers. So hope you understand the relationship between how Spearman's rho is computing the correlation in outliers. Also, the Pearson's correlation coefficient is computing concerned outliers when they compute the r value.

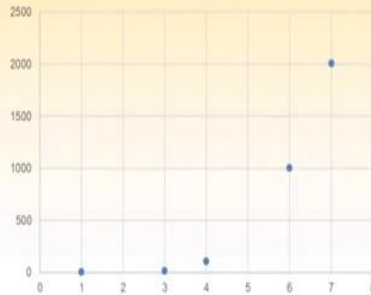
(Refer Slide Time: 11:25)

Activity

Spearman's rank Correlation

What will be r and ρ values?

X	Y
1	1
3	10
4	100
6	1000
7	2000

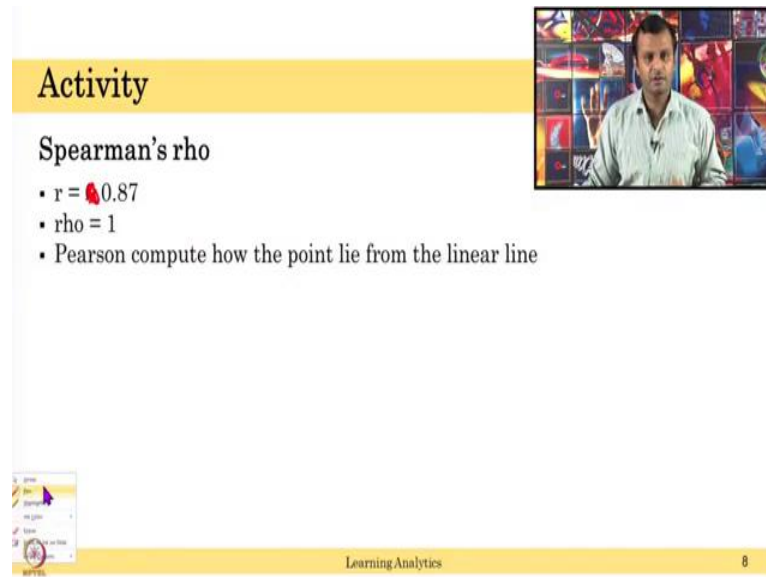


Let us do the other activity to make it clear. You see X value in this plot is 1, 3, 4, 6, 7 or some scale between 1 to 10 but Y value is 1, 10, 100, 1000, 2000 like in the math log scale or something. If it is 10000 that means the logarithmic scale. And if I want to compute Spearman's rank correlation and also Pearson's correlation coefficient.

What will be the values? You do not need to do computations, instead based on your understanding till now, try to find out the correlation coefficient between these two variables. .

Can you guess the ρ and r value after doing that please resume to continue.

(Refer Slide Time: 12:32)



Activity

Spearman's rho

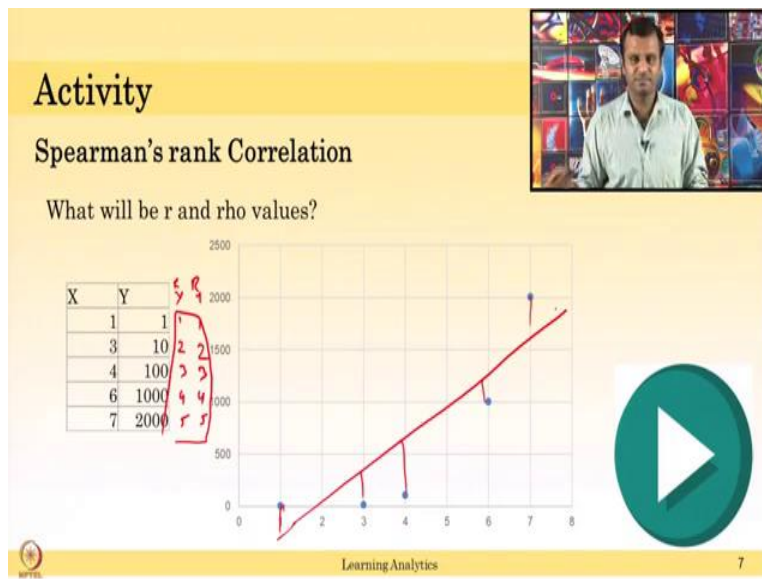
- $r = 0.87$
- $\rho = 1$
- Pearson compute how the point lie from the linear line

Learning Analytics 8

The r value is 0.87 positive and it indicates it is highly correlated but some points are not linearly related with X and Y . If X increases Y also increases but it is not fitting in the linear relationship. But ρ indicates 1, it means for all the incremental in the X , Y also increases they may not be in the linear scale of increment, but there is an increment in X , there is Y is incremented.

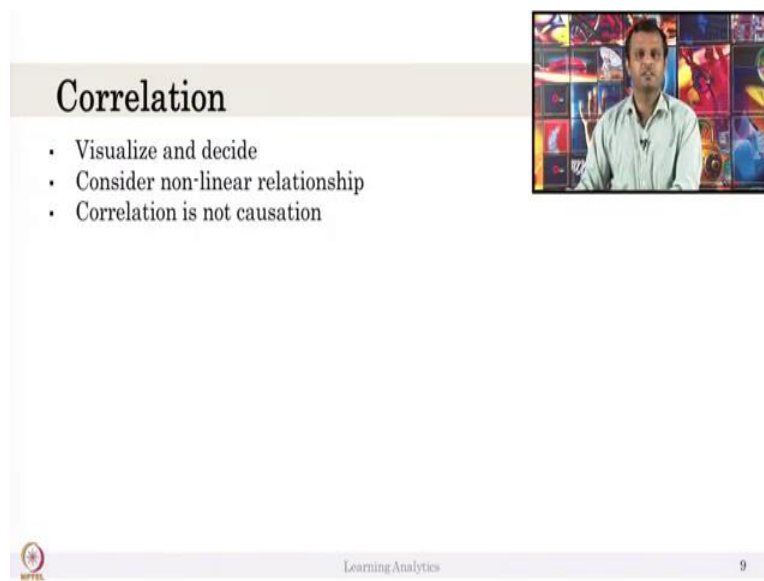
So, value, 1 indicates whenever the X increases Y also increases, maybe the different scale but there is a relationship. So this indicates ρ can handle, if one variable in a different scale that is logarithm scale and X in other scale, it can even be able to identify the correlation coefficient because we add a rank. So, all the ranks are same, if find a difference it will be 0. So you get 1 minus 0, it will be 1 that is how ρ is computed but, the Pearson's correlation coefficient will try to indicate 0.87.

(Refer Slide Time: 14:13)



In the previous slide, we have seen, if you just compute the rank, rank of X will be 1, 2, 3, 4, 5 and rank of Y will be 1, 2, 3, 4, 5. So if you find a difference between these two values definitely just 0. And by using Spearman formula, we have $1 - 0$ divided by large value, its value will be 1. So, that is why the Spearman's rho is 1 but the r value is not 1, as r value is trying to fit a for example line line kind of kind of fit like this. There is a small difference between these lines values in these lines. That is why it is 0.81.

(Refer Slide Time: 15:02)



The slide is titled "Correlation" in a large, bold, black font. Below the title, there is a bulleted list with three items: "Visualize and decide", "Consider non-linear relationship", and "Correlation is not causation". In the top right corner, there is a small video inset showing a man with dark hair, wearing a light blue shirt, speaking. The background of the slide is white. At the bottom left, there is a small circular logo with a star. At the bottom center, the text "Learning Analytics" is visible. At the bottom right, the number "9" is displayed.

- Visualize and decide
- Consider non-linear relationship
- Correlation is not causation

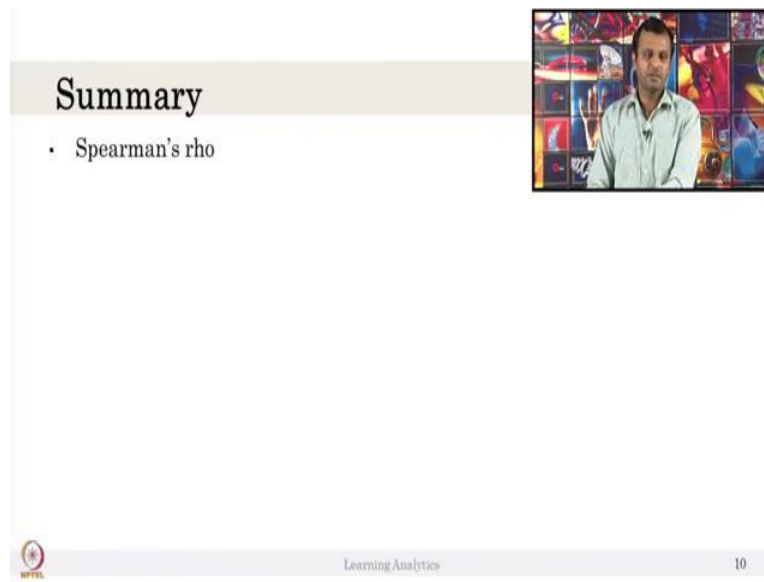
Hope you understand what is Spearman's rho and correlation coefficient and Pearson's correlation coefficient are. So, which one to choose is very very important and can I just go and use coefficient as the values to prove algorithm or hypothesis or my influences. I would like to say it is not the case. So, first always visualize your data as visual a figure diagnostic analysis means, what are the relationship between X and Y? Is X and Y in the same scale or linearity all these things then you can choose which correlation coefficient you want to be put.

Most widely Pearson's correlation coefficient is used if both values are nominal. If one of the values is ordinal, people Spearman's rho.

So, look at your data or visualize since you know the math about both of these correlations Spearman's and Pearson's. Now, you know which one to choose based on your data. So visualize it and decide and also consider non-linear relationship between this data then pick the right a correlation coefficient.

Most importantantly, correlation is not causation. Correlation indicates just what is the relationship between X and Y. If X increases is Y also increasing or decreasing that is what the correlation is indicating. Never use correlation as a causation. It never tells that Y increases because of X increases. We say there is a correlation between these two variables, but it never tells causation of Y. why is increasing? So never use the correlation coefficient values to prove your causation in our theory.

(Refer Slide Time: 17:00)



Summary

- Spearman's rho

Learning Analytics 10

So, in this video we saw what is Spearman's rho and we also saw what is relationship between Spearman's rho and Pearson's correlation coefficient. Hope you understood both. Thank you.