# Marketing Research and Analysis-II

Prof. Dr Jogendra Kumar Nayak
Department of Management Studies
Indian Institute of Technology Roorkee

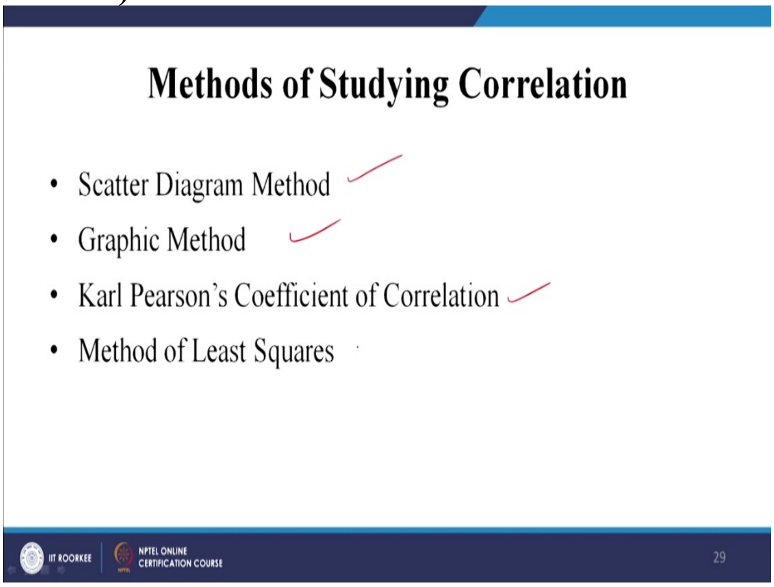
## Lecture – 42
## Correlation in SPSS

Hello friends. Welcome to the class of marketing research and analysis. In the last lecture we had started with the concept of covariance and correlation where I explained you what is basically the utility the use and the meaning of covariance and correlation how correlation is much powerful technique over the covariance the reason behind it and I said that covariance is slightly becomes awkward due to the units that it uses of for 2 variables.

That means 2 variables could be having very different units and that sometimes looks very difficult very absurd to compare. And that is why we need to have another technique which can standardize them and give us a solution so that it is easier for us to compare. So covariance was basically telling us about the direction and not the degree or intensity where as correlation on the other hand tells us both the degree and the direction. So that is the advantage.

So we will proceed with today's class with the correlation. So how do you know study correlation, so basically from a visual point of view.

**(Refer Slide Time: 01:42)**



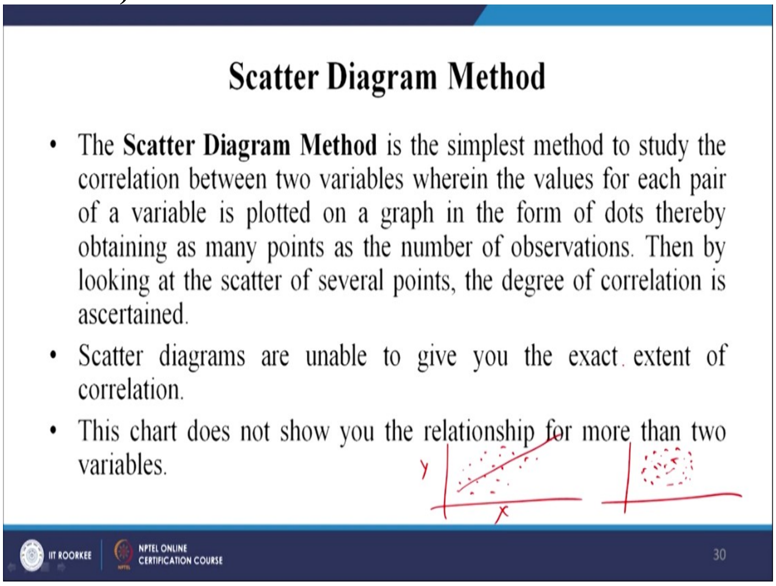
Methods of Studying Correlation

- Scatter Diagram Method
- Graphic Method
- Karl Pearson's Coefficient of Correlation
- Method of Least Squares

IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE
29

You can have a Scatter Diagram Method or a Graphic Method or we have a Karl Pearson's Coefficient of correlation of course it and Method of Least Squares. So these are the ways you know used to measure the correlation or check the correlation.

**(Refer Slide Time: 01:52)**



## Scatter Diagram Method

- The **Scatter Diagram Method** is the simplest method to study the correlation between two variables wherein the values for each pair of a variable is plotted on a graph in the form of dots thereby obtaining as many points as the number of observations. Then by looking at the scatter of several points, the degree of correlation is ascertained.
- Scatter diagrams are unable to give you the exact extent of correlation.
- This chart does not show you the relationship for more than two variables.

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE

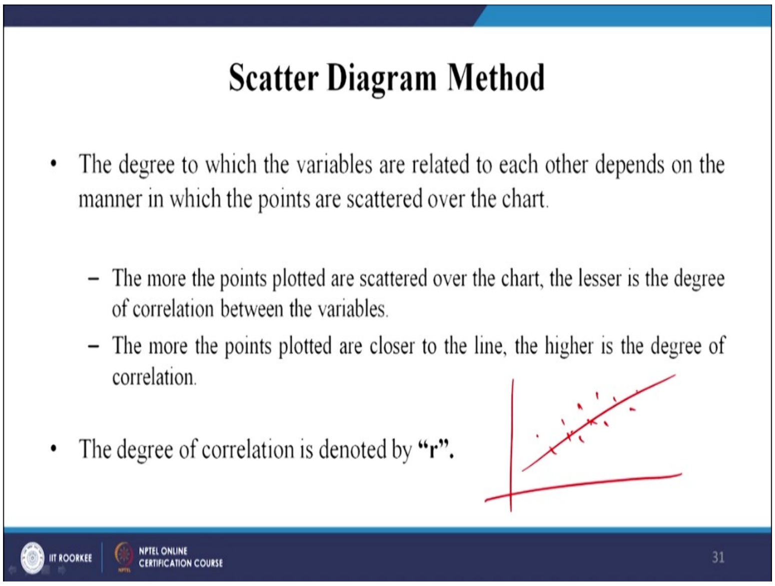
30

Now what is the Scatter Diagram? Scatter Diagram Method is the very simple method to study the correlation between 2 variables wherein the values for each pair is plotted on a graph in the form of dots thereby obtaining as many points as the number of observations. Then by looking at the Scatter of the several points the degree of correlation is ascertained that means it looks something like this.

For example, if I plot certain dots for example these are the dots. So dots are each my observations. So, from here let us say this is my X and this is my Y. So can I say what kind of correlationship exists? So if I draw a line in between so I can say well there is some positive correlation it is increasing it seems. But suppose I have another relationship something like this.

So here if I draw what is a correlation ship it is very difficult to ascertain because it is in fact that I can hardly say there is any relationship? So scattered diagrams help you to visually understand that but there are unable to give you the exact extent of correlation how much the which is the most important part of correlation is the how much that; how much is the what you do not get here. This chart does not show you the relationship for more than 2 variables only 2 variables x and y but not if there are 3 variables multiple correlation you cannot do it.

**(Refer Slide Time: 03:18)**

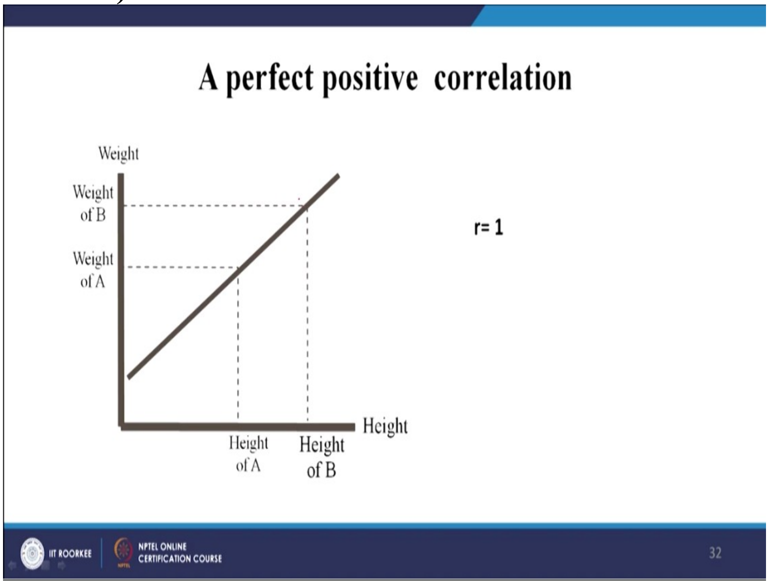


## Scatter Diagram Method

- The degree to which the variables are related to each other depends on the manner in which the points are scattered over the chart.
  - The more the points plotted are scattered over the chart, the lesser is the degree of correlation between the variables.
  - The more the points plotted are closer to the line, the higher is the degree of correlation.
- The degree of correlation is denoted by **"r".**

The degree to which the variables are related to each other depends on the manner in which the points are scattered over the chart. Yes, that is true. How they are scattered, is there any pattern or not that decides. The more the points plotted are scattered over the chart, the less is the degree of correlation that means is the spread is too large and it is in a systematic pattern then the correlation is weak the more the points plotted are closer to the line that line which I said suppose this is the line right the higher is the degree of the correlation so the closer they are to the line the higher is the degree of correlation between the 2 variables it is denoted by r.

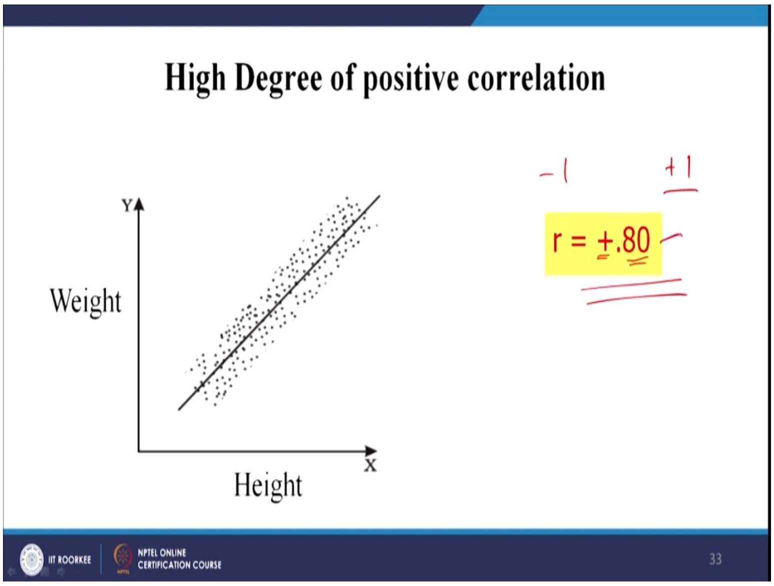
**(Refer Slide Time: 04:00)**



## A perfect positive correlation

So this is then you see the perfect correlation positive correlation where r = 1 so r =1 weight and we are taken height so when we are saying it is a perfect positive correlation.

**(Refer Slide Time: 04:13)**



High Degree of positive correlation

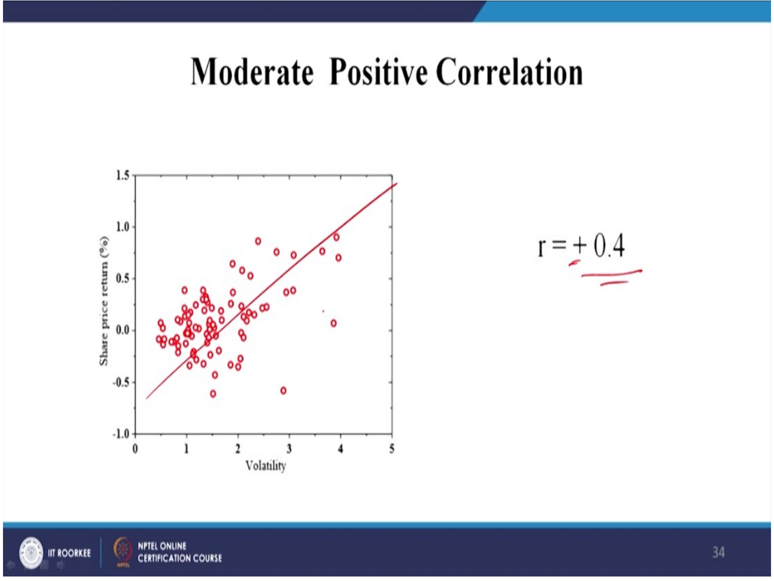
-1 +1

r = +.80

IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 33

When r = .8 when we are measured r we will see how to measure r. But when height and weight we are saying r = .8. So we say it is a substantially strong positive correlation it is a positive because of the sign and because of the value lies between -1 to +1. So it is close to +1 so highly positive.

**(Refer Slide Time: 04:35)**



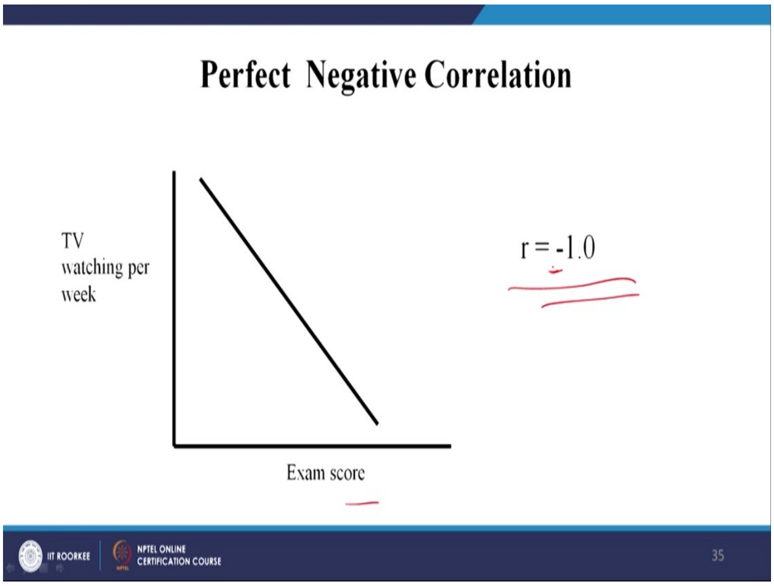
Moderate Positive Correlation

r = + 0.4

IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 34

Moderately positive now, so this is somewhere sign is positive and 0.4. If I draw a line you see. So this looks little awkward but still it is a positive correlation.

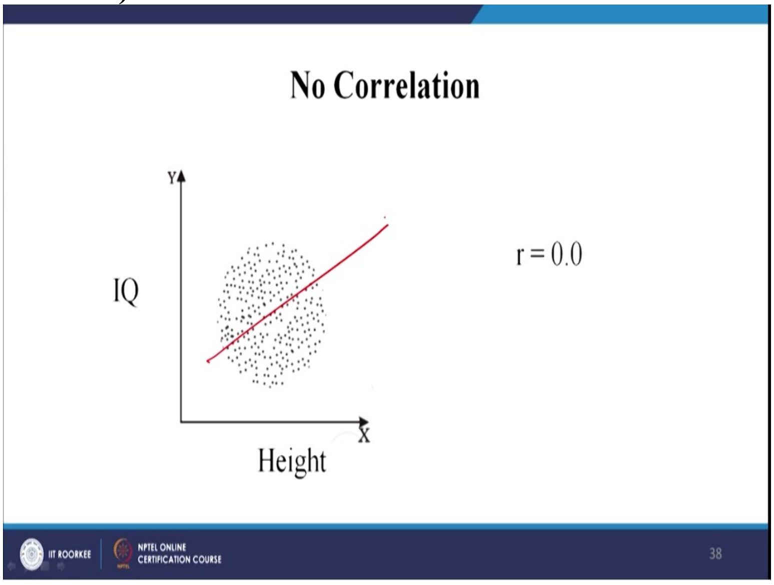
**(Refer Slide Time: 04:47)**



## Perfect Negative Correlation

Now this is a Perfect Negative you see TV watching per week and scores in the exam. So, my relationship is negative as TV watching is high. My marks are low. As a TV watching goes down my marks are increasing a Perfect Correlation. But the direction is inverse Negative relationship that means if one increases the other decreases, moderately negative. Moderately negative similarly, so -0.8. No correlation.

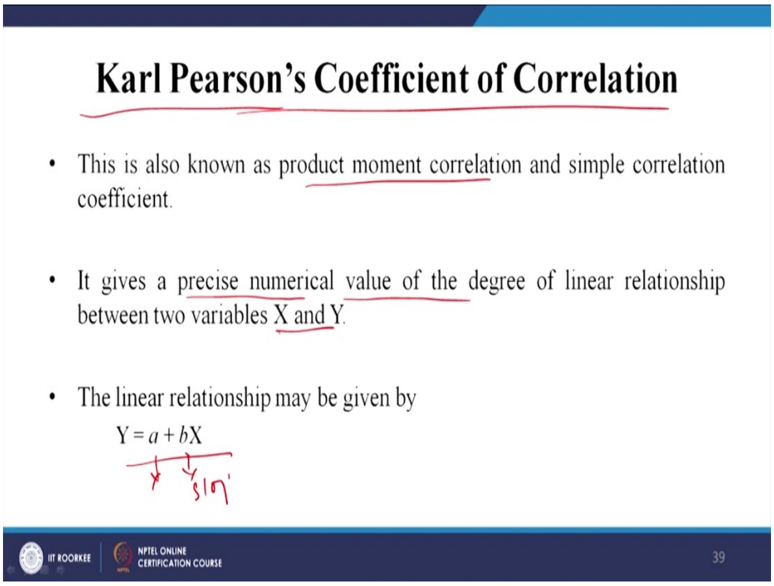
**(Refer Slide Time: 05:13)**



## No Correlation

So here you see IQ and height. Is there any relationship between IQ and height .Well if I draw what is the correlation coming nothing.

**(Refer Slide Time: 05:21)**



## Karl Pearson's Coefficient of Correlation

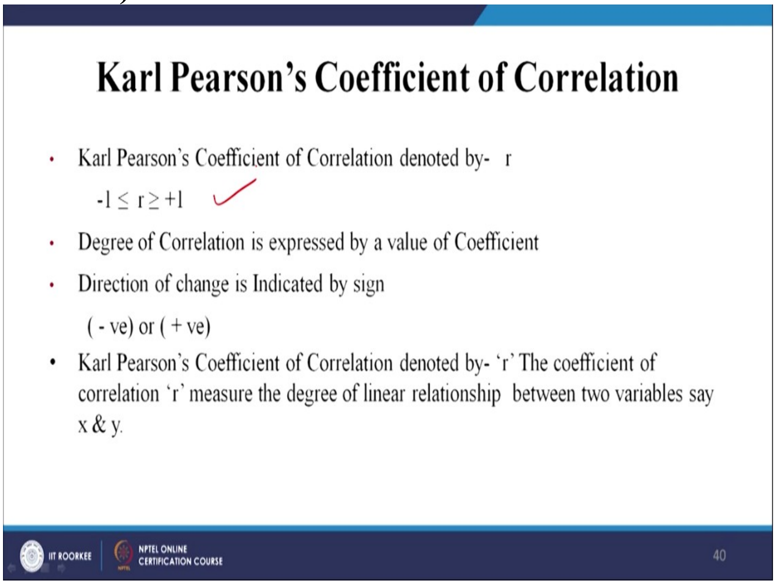
- This is also known as product moment correlation and simple correlation coefficient.
- It gives a precise numerical value of the degree of linear relationship between two variables X and Y.
- The linear relationship may be given by

  $Y = a + bX$

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 39

So the best method of measuring statistically is called the Karl Pearson's method of Coefficient of correlation. So this is also called as product moment correlation. You must have seen somewhere written as simple correlation or product moment correlation. It gives a precise numerical value of the degree of linear relationship between 2 variables X and Y. So what is the degree of the relationship and this relationship may be given by a regression equation which is Y = a + b of X, where Y is my intercept, B is my slope and X is my independent variable.

**(Refer Slide Time: 05:56)**



## Karl Pearson's Coefficient of Correlation

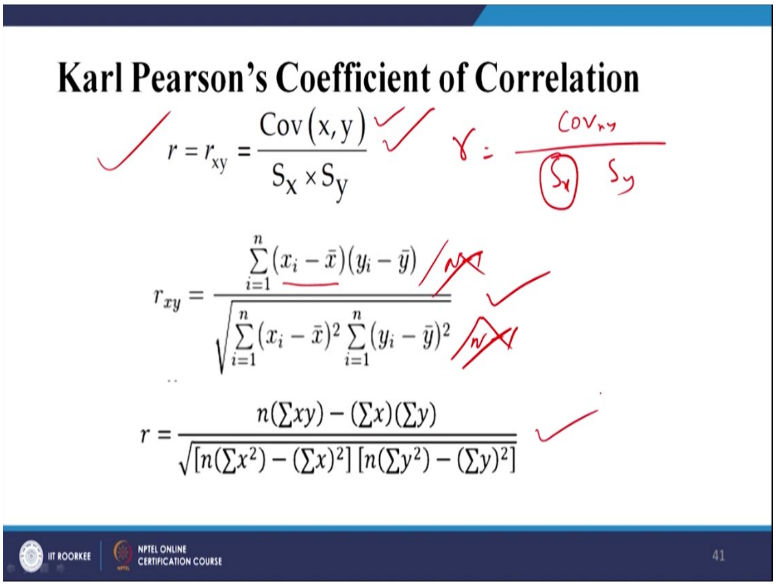
- Karl Pearson's Coefficient of Correlation denoted by- r

  $-1 \leq r \geq +1$ ✓
- Degree of Correlation is expressed by a value of Coefficient
- Direction of change is Indicated by sign

  ( - ve) or ( + ve)
- Karl Pearson's Coefficient of Correlation denoted by- 'r' The coefficient of correlation 'r' measure the degree of linear relationship between two variables say x & y.

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 40

**(Refer Slide Time: 06:00)**



## Karl Pearson's Coefficient of Correlation

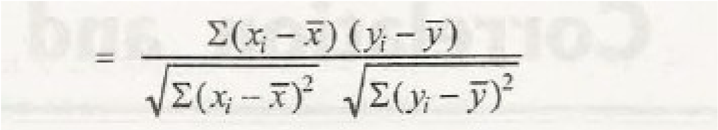
$$r = r_{xy} = \frac{\mathrm{Cov}(x,y)}{S_x \times S_y} \qquad r = \frac{Cov_{xy}}{\boxed{S_x}\; S_y}$$

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) / \cancel{n-1}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2} / \cancel{n-1}}$$

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2]\,[n(\sum y^2) - (\sum y)^2]}}$$

IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 41

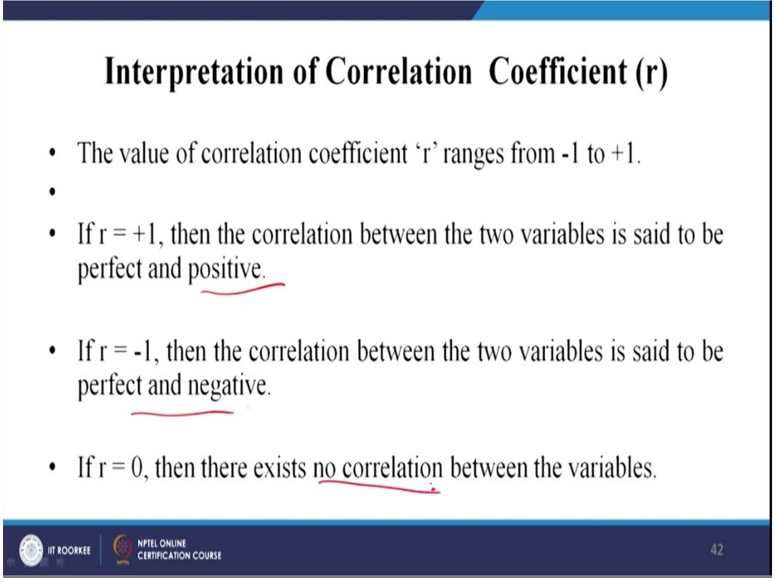
As I said it lies between -1 to +1. So this is the formula, so how do you measure the Karl Pearson Coefficient of Correlation. The formula is very simple. You need to understand. So what we said correlation is a standardized form of the covariance. So if my covariance value, if I divide my covariance value between x and y with the standard deviation of x and standard deviation y. This is what it becomes as the; my coefficient of correlation.



$$= \frac{\Sigma(x_i-\bar{x})\,(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2}\;\sqrt{\Sigma(y_i-\bar{y})^2}}$$

The N -1 here and the N-1 here actually are cutting each other. So that is why we are not showing it is not required anymore. So this is the way you write. This is a very simple thing. You must remember this much and nothing you do not have to remember any other thing. This is another way of representing the same thing.

**(Refer Slide Time: 07:16)**



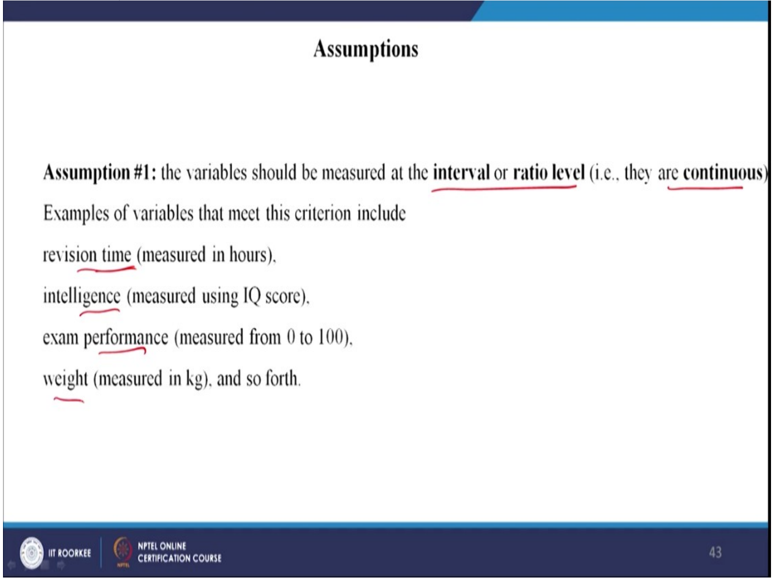
## Interpretation of Correlation Coefficient (r)

- The value of correlation coefficient 'r' ranges from -1 to +1.
- 
- If r = +1, then the correlation between the two variables is said to be perfect and positive.
- If r = -1, then the correlation between the two variables is said to be perfect and negative.
- If r = 0, then there exists no correlation between the variables.

IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 42

So let us see, so interpretation of the correlation coefficient. The value this is said, if it is +1 perfect and positive, if it is - 1 perfect and negative. If r = 0, no correlation.

**(Refer Slide Time: 07:29)**



### Assumptions

**Assumption #1:** the variables should be measured at the **interval** or **ratio level** (i.e., they are **continuous**)

Examples of variables that meet this criterion include

revision time (measured in hours),

intelligence (measured using IQ score),

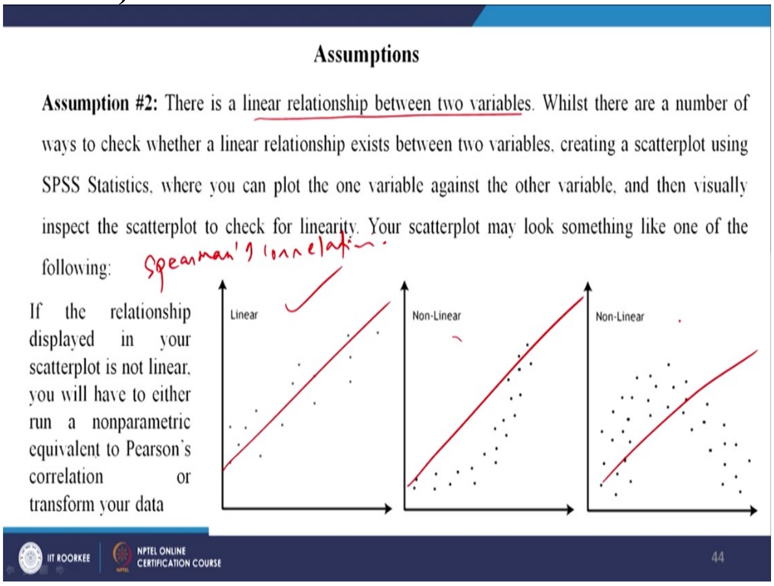
exam performance (measured from 0 to 100),

weight (measured in kg), and so forth.

IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE
43

Some assumptions in correlation what are the assumptions? The variables should be measured at the interval or ratio level. First they should be continuous for example time revision time, intelligence, exam performance, weight.

**(Refer Slide Time: 07:46)**



### Assumptions

**Assumption #2:** There is a linear relationship between two variables. Whilst there are a number of ways to check whether a linear relationship exists between two variables, creating a scatterplot using SPSS Statistics, where you can plot the one variable against the other variable, and then visually inspect the scatterplot to check for linearity. Your scatterplot may look something like one of the following:

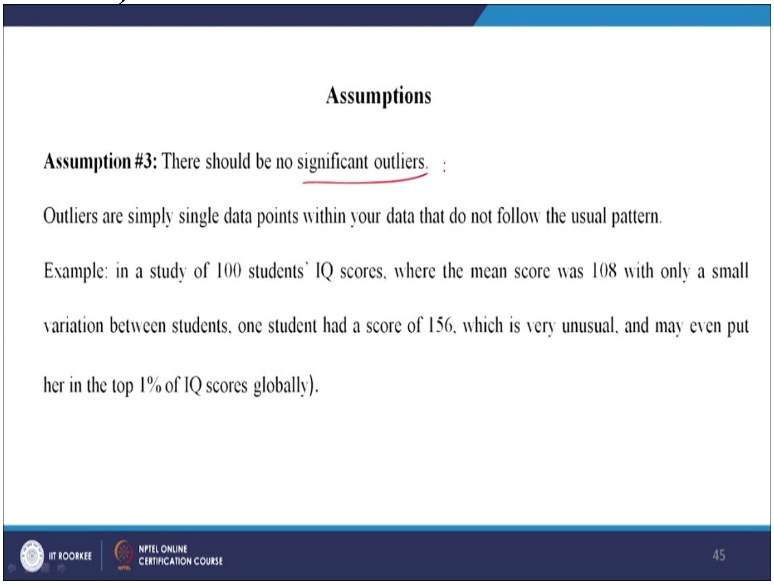
Spearman's correlation.

If the relationship displayed in your scatterplot is not linear, you will have to either run a nonparametric equivalent to Pearson's correlation or transform your data

There is a linear relationship between the 2. The assumption this is an assumption in correlation that the relationship is linear. Well there are a number of ways to check a linear relationship creating a scatter plot I guess each other and then visually inspecting is the simplest way. So you see if I draw a line you see I draw a line and here also I draw a line. So, these 2 are nonlinear but this is linear. If the relationship displayed is not linear. You will have to either run in non-parametric equivalent to the Person's correlation, what is the non parametric equivalent? I have explained it earlier also when I was teaching non parametric techniques.

In which I said there is a technique called Spearman's correlation which is used as a non-parametric test for the correlation. So or you have to transform your data and convert it from non normal into a normally distributed data. If it is non normal you have to convert into normally distributed data. If data is normally distributed the chances are very fair that it will become following a linear pattern.

**(Refer Slide Time: 09:04)**



Assumptions

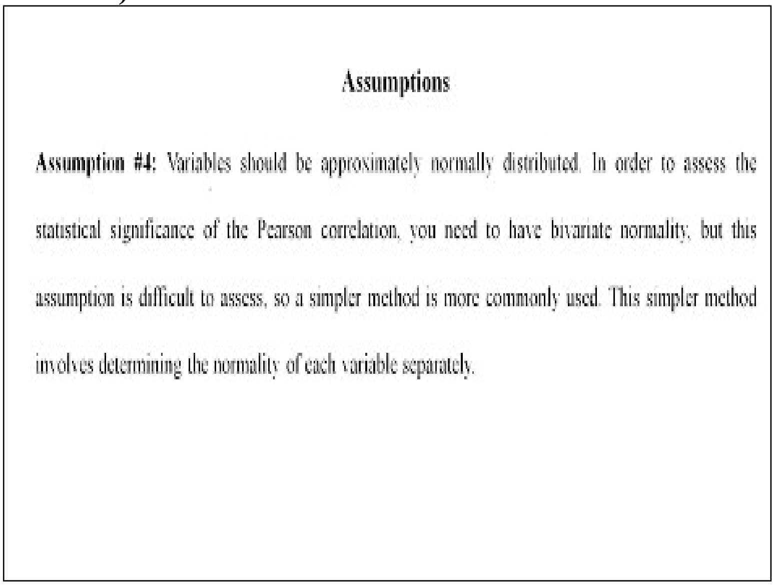
**Assumption #3:** There should be no significant outliers.

Outliers are simply single data points within your data that do not follow the usual pattern.

Example: in a study of 100 students' IQ scores, where the mean score was 108 with only a small variation between students, one student had a score of 156, which is very unusual, and may even put her in the top 1% of IQ scores globally).

The third assumption is no significant outlier should be there. If there is an outlier it will distort the entire situation.

**(Refer Slide Time: 09:06)**



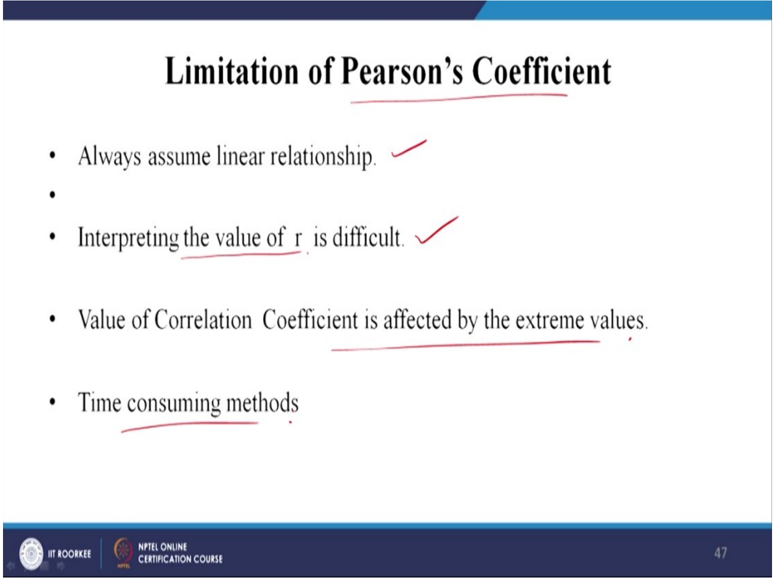
Assumptions

**Assumption #4:** Variables should be approximately normally distributed. In order to assess the statistical significance of the Pearson correlation, you need to have bivariate normality, but this assumption is difficult to assess, so a simpler method is more commonly used. This simpler method involves determining the normality of each variable separately.

Fourth the variables should be approximately normally distributed in order to assess the statistical significance of the correlation. Now what does it mean. If you do not have understanding whether it is statistically significant or not between the relationships between 2 variables, then although you get a degree you cannot say it will happen every time the same way.

Maybe it has happened this time by chance but maybe another time it will not give the same result, so but if it is significant that means it is not by chance it will happen again and again and again in the same way. So but this method is difficult to assess. So, simpler common method is used. This method involves determining the normality of each variable separately. So what it says is you do not check it at the same time but individually you check whether the data or the variables are normally distributed or not.

**(Refer Slide Time: 10:02)**



## Limitation of Pearson's Coefficient

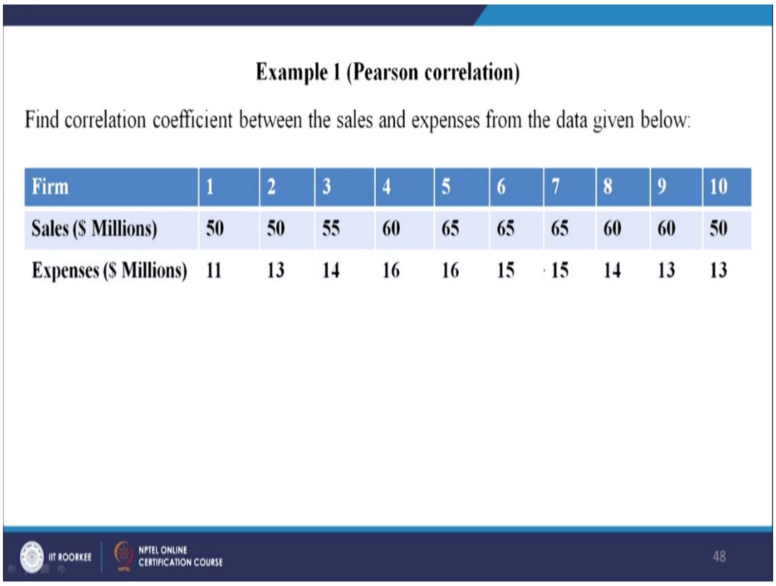
- Always assume linear relationship. ✓
- 
- Interpreting the value of r is difficult. ✓
- Value of Correlation Coefficient is affected by the extreme values.
- Time consuming methods

IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE
47

What are the limitations of Pearson's coefficient? Does it have any limitation? Yes. First of all, assuming in linear relationship in life I said earlier key most of the relationship may not be linear they could be non-linear. Interpreting the value of r is sometimes difficult although we say if it is +1 it is extremely positive -1 extremely negative but it is still interpretation is little challenging.

There effected by the extreme values outliers and there it is time consuming to calculate and understand because in real life there is nothing comes in a plate to you. So suppose the situation is given and then you have to interpret what are the variables and then find out what the relationship is not that simple as it looks when I give you on a pen and paper.

**(Refer Slide Time: 10:51)**



### Example 1 (Pearson correlation)

Find correlation coefficient between the sales and expenses from the data given below:

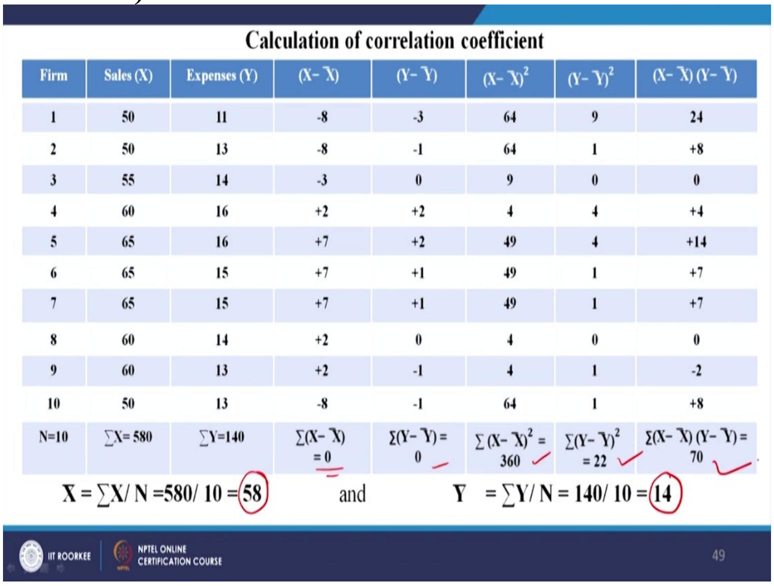
| Firm | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Sales ($ Millions) | 50 | 50 | 55 | 60 | 65 | 65 | 65 | 60 | 60 | 50 |
| Expenses ($ Millions) | 11 | 13 | 14 | 16 | 16 | 15 | 15 | 14 | 13 | 13 |

IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 48

Now let us solve a problem find the correlation coefficient between the sales and expenses from the data given below the sales is given 50 number of maybe you know different Firms are there. Firm 1 to Firm 10 and the sales is given 50, 50, 55, 60, 65 and goes on. This is all in Millions. The expenses conducted by this or expanded by these companies are 11, 13, 14, 16 to something.

Now I am interested to know what is the direction of the relationship. Is it a positive relationship or negative relationship or there is no relationship? And what is the degree of intensity of the relationship how strong is the relationship. So to do this we if we cannot we will not do covariance, we will do a correlation in this case and luckily these are having same units also. So let us say.

**(Refer Slide Time: 11:43)**



### Calculation of correlation coefficient

| Firm | Sales (X) | Expenses (Y) | $(X-\bar{X})$ | $(Y-\bar{Y})$ | $(X-\bar{X})^2$ | $(Y-\bar{Y})^2$ | $(X-\bar{X})(Y-\bar{Y})$ |
|---|---|---|---|---|---|---|---|
| 1 | 50 | 11 | -8 | -3 | 64 | 9 | 24 |
| 2 | 50 | 13 | -8 | -1 | 64 | 1 | +8 |
| 3 | 55 | 14 | -3 | 0 | 9 | 0 | 0 |
| 4 | 60 | 16 | +2 | +2 | 4 | 4 | +4 |
| 5 | 65 | 16 | +7 | +2 | 49 | 4 | +14 |
| 6 | 65 | 15 | +7 | +1 | 49 | 1 | +7 |
| 7 | 65 | 15 | +7 | +1 | 49 | 1 | +7 |
| 8 | 60 | 14 | +2 | 0 | 4 | 0 | 0 |
| 9 | 60 | 13 | +2 | -1 | 4 | 1 | -2 |
| 10 | 50 | 13 | -8 | -1 | 64 | 1 | +8 |
| N=10 | $\sum X = 580$ | $\sum Y=140$ | $\sum(X-\bar{X}) = 0$ | $\sum(Y-\bar{Y}) = 0$ | $\sum(X-\bar{X})^2 = 360$ ✓ | $\sum(Y-\bar{Y})^2 = 22$ ✓ | $\sum(X-\bar{X})(Y-\bar{Y}) = 70$ ✓ |

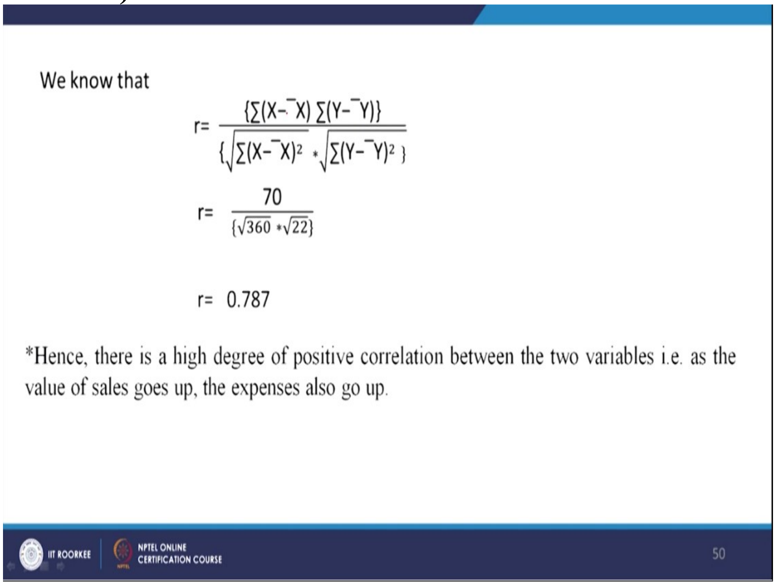
$\bar{X} = \sum X/ N = 580/ 10 = \boxed{58}$ and $\bar{Y} = \sum Y/ N = 140/ 10 = \boxed{14}$

IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 49

How do we go about it. So these are my firms. So the 10 Firms sales is given. So my summation of the sales is 580, summation of my expenses 140. Now I need $(X – \bar{X})$ . So what is the formula covariance by standard deviation of X into standard deviation of Y. So first we find out $(X – \bar{X})$, so $\bar{X}$ is how much 580 / 10. So this is my $\bar{X}$ , $\bar{y}$ is 14, so $(X – \bar{X})$ - 8 ,– 8, - 3 . It goes on.

So, this totally 0, similarly, if you look at $(Y – \bar{y})$ is again 0 so, because they adjust each other but for that we will take the X - X square now. So, what is the $(X – \bar{X})^2$ now 360 and $(Y – \bar{y})^2$ is 22. Now the product deviation of $(X – \bar{X})* (Y – \bar{y})$, this is equal to if you multiply 8 x3= 24, 8 * 1= 8, - 3 * 0 = 0 so goes on 70.

**(Refer Slide Time: 13:03)**



We know that

$$r= \frac{\{\sum(X-\bar{X}) \sum(Y-\bar{Y})\}}{\{\sqrt{\sum(X-\bar{X})^2} * \sqrt{\sum(Y-\bar{Y})^2}\}}$$

$$r= \frac{70}{\{\sqrt{360} * \sqrt{22}\}}$$

$$r= 0.787$$

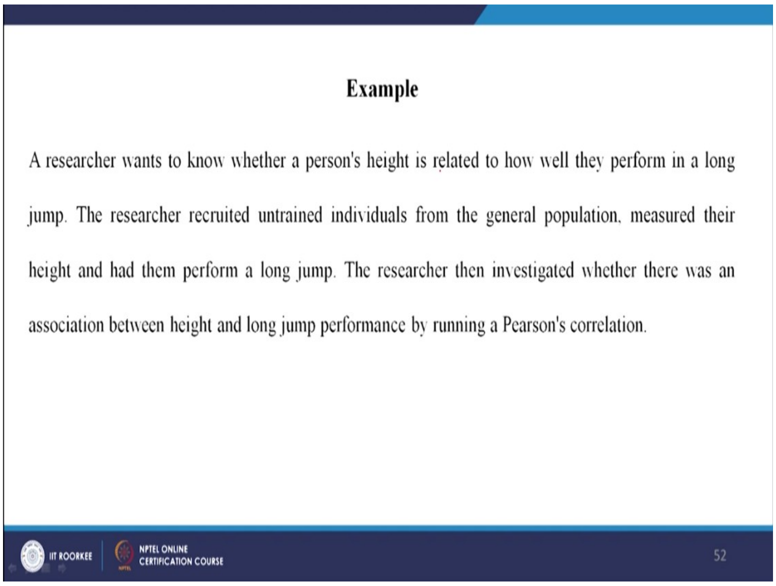
*Hence, there is a high degree of positive correlation between the two variables i.e. as the value of sales goes up, the expenses also go up.

IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE
50

So what is the formula $(X – \bar{X})$, how much it has come let us check. So $(X – \bar{X})* (Y – \bar{y})$ so this is 70 divided by $(X – \bar{X})^2$, so $(X – \bar{X})^2$ is 360 and this is 22. $\sqrt{360} * \sqrt{22}$. So this is equal to 0.787. So, my correlation between the sales and expense is first of all. Is it positive or negative. It is positive. And that means if sales increases expenses; if expenses increase my sales increases. And it is 787 means it is close to 1.

So it is the high. So there is a high degree of positive correlation between the 2 variables that is as this value of sales goes up the expenses also go up.

**(Refer Slide Time: 13:56)**

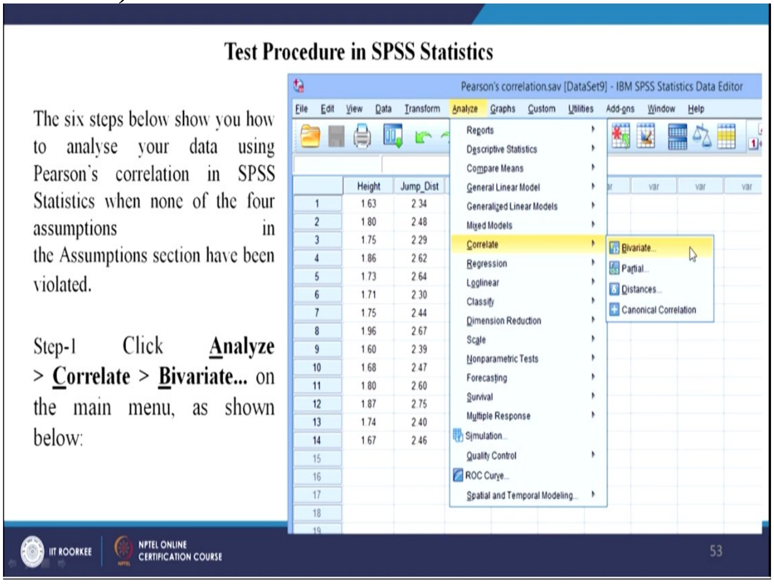


## Example

A researcher wants to know whether a person's height is related to how well they perform in a long jump. The researcher recruited untrained individuals from the general population, measured their height and had them perform a long jump. The researcher then investigated whether there was an association between height and long jump performance by running a Pearson's correlation.

Now I will show you how to do this in the SPSS. Now this is a case where a researcher wants to know whether a person's height. This is an example I have taken but I will use a separate example.

**(Refer Slide Time: 14:06)**



## Test Procedure in SPSS Statistics

The six steps below show you how to analyse your data using Pearson's correlation in SPSS Statistics when none of the four assumptions in the Assumptions section have been violated.

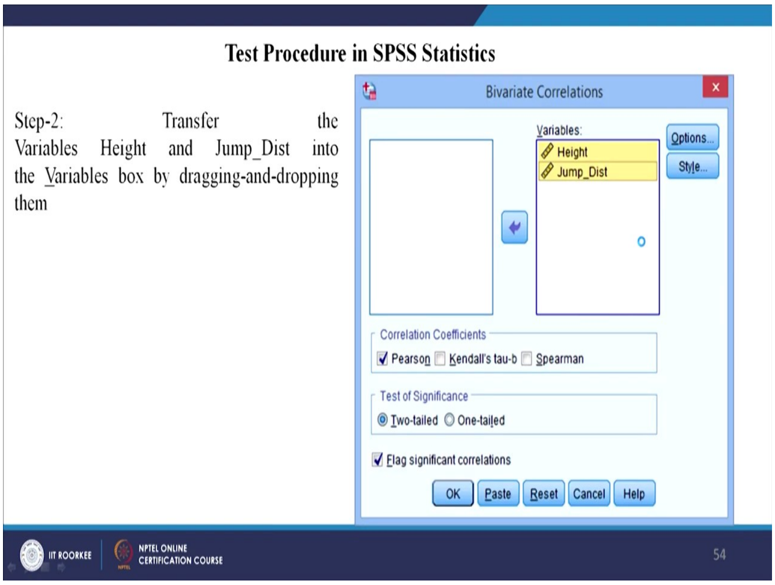
Step-1 Click **Analyze** > **Correlate** > **Bivariate...** on the main menu, as shown below:

Pearson's correlation.sav [DataSet9] - IBM SPSS Statistics Data Editor

|  | Height | Jump_Dist |
|---|---|---|
| 1 | 1.63 | 2.34 |
| 2 | 1.80 | 2.48 |
| 3 | 1.75 | 2.29 |
| 4 | 1.86 | 2.62 |
| 5 | 1.73 | 2.64 |
| 6 | 1.71 | 2.30 |
| 7 | 1.75 | 2.44 |
| 8 | 1.96 | 2.67 |
| 9 | 1.60 | 2.39 |
| 10 | 1.68 | 2.47 |
| 11 | 1.80 | 2.60 |
| 12 | 1.87 | 2.75 |
| 13 | 1.74 | 2.40 |
| 14 | 1.67 | 2.46 |

In this case just to show you how to what is the steps. So what you do is you go to analyze go to correlate and go to bivariate.

**(Refer Slide Time: 14:16)**



Test Procedure in SPSS Statistics

Step-2: Transfer the Variables Height and Jump_Dist into the Variables box by dragging-and-dropping them

Bivariate Correlations

Variables: Height, Jump_Dist

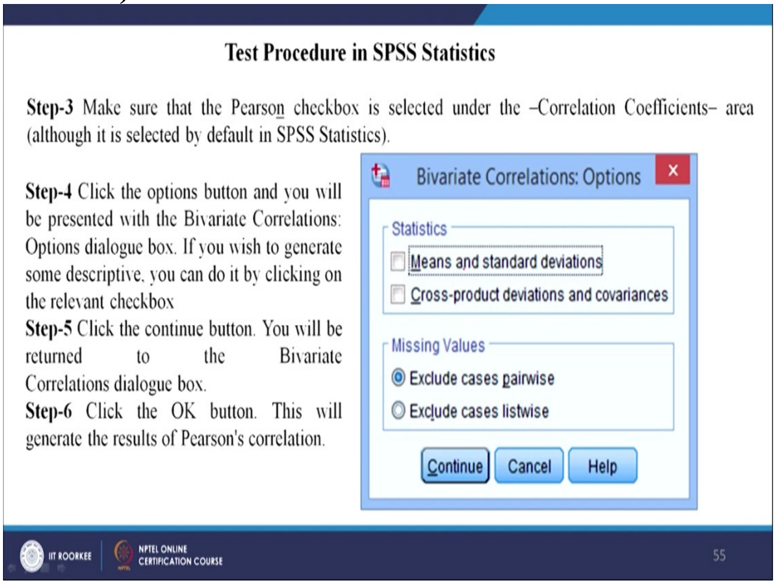
Correlation Coefficients: ☑ Pearson □ Kendall's tau-b □ Spearman

Test of Significance: ● Two-tailed ○ One-tailed

☑ Flag significant correlations

After taking going to bivariate you take the bivariate why it is called bivariate because it is a 2 variables Bi 2, Bicycle 2 wheels. So, height and suppose the Jumping distance in this case it was; in this example. So and what is the tail and you have to take.

**(Refer Slide Time: 14:35)**



Test Procedure in SPSS Statistics

**Step-3** Make sure that the Pearson checkbox is selected under the –Correlation Coefficients– area (although it is selected by default in SPSS Statistics).

**Step-4** Click the options button and you will be presented with the Bivariate Correlations: Options dialogue box. If you wish to generate some descriptive, you can do it by clicking on the relevant checkbox

**Step-5** Click the continue button. You will be returned to the Bivariate Correlations dialogue box.

**Step-6** Click the OK button. This will generate the results of Pearson's correlation.

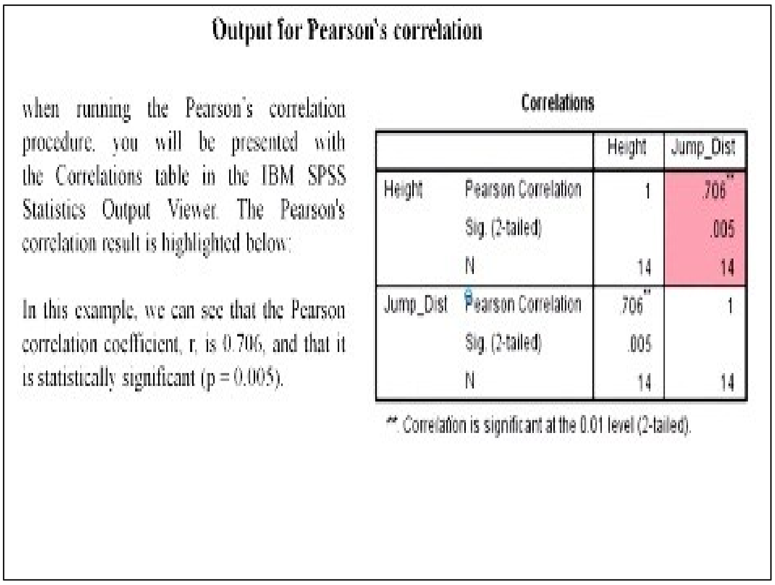
Bivariate Correlations: Options

Statistics: □ Means and standard deviations □ Cross-product deviations and covariances

Missing Values: ● Exclude cases pairwise ○ Exclude cases listwise

So then you have to go for this and you have to seek you whether you require the mean and standard deviation are not all these things.

**(Refer Slide Time: 14:43)**



## Output for Pearson's correlation

when running the Pearson's correlation procedure, you will be presented with the Correlations table in the IBM SPSS Statistics Output Viewer. The Pearson's correlation result is highlighted below:

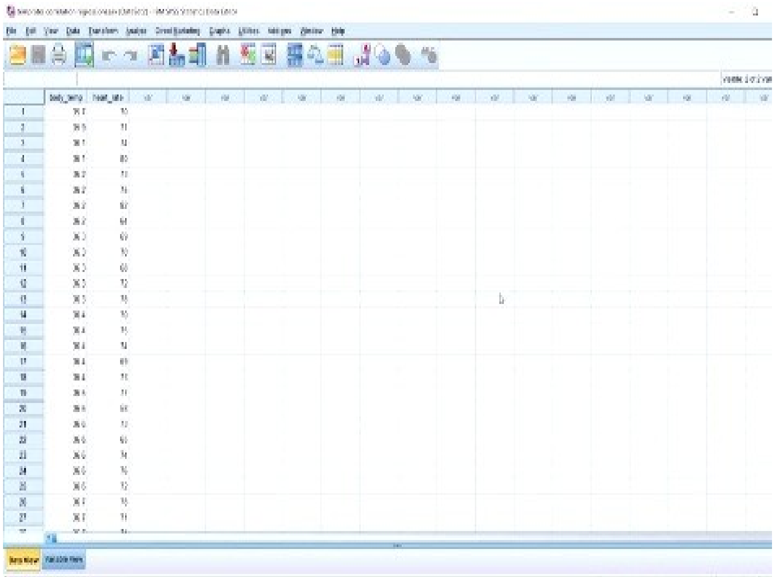
In this example, we can see that the Pearson correlation coefficient, r, is 0.706, and that it is statistically significant (p = 0.005).

**Correlations**

| | | Height | Jump_Dist |
|---|---|---|---|
| Height | Pearson Correlation | 1 | .706** |
| | Sig. (2-tailed) | | .005 |
| | N | 14 | 14 |
| Jump_Dist | Pearson Correlation | .706** | 1 |
| | Sig. (2-tailed) | .005 | |
| | N | 14 | 14 |

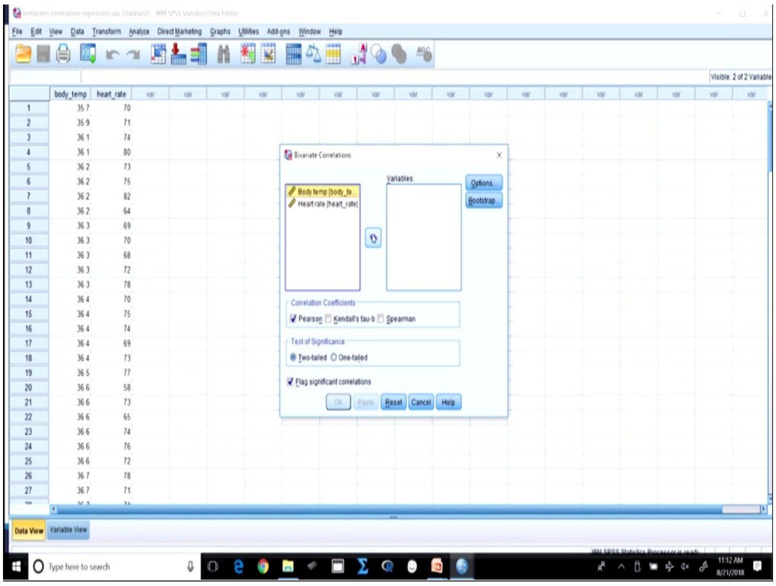
**. Correlation is significant at the 0.01 level (2-tailed).

**(Refer Slide Time: 14:49)**


File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
body_temp
heart_rate
Data View
Variable View

And let me show you here first let me show you then will try to analyze. So let me take this case, now what I have done is this case is a case where we have taken values of the correlation we want to find out the correlation between body temperature and the heart rate the heartbeat rate. So these are some other values for another 120 people I think. So, 130 people in fact. So if I want to see what is the correlation between is there any association between body temperature and heart rate. I want to check that. So to do that how do I go into SPSS now first I will go to analyze I will go to correlate and then go to Bivariate.

**(Refer Slide Time: 15:29)**



| | body_temp | heart_rate |
|---|---|---|
| 1 | 35.7 | 70 |
| 2 | 35.9 | 71 |
| 3 | 36.1 | 74 |
| 4 | 36.1 | 80 |
| 5 | 36.2 | 73 |
| 6 | 36.2 | 75 |
| 7 | 36.2 | 82 |
| 8 | 36.2 | 64 |
| 9 | 36.3 | 69 |
| 10 | 36.3 | 70 |
| 11 | 36.3 | 68 |
| 12 | 36.3 | 72 |
| 13 | 36.3 | 78 |
| 14 | 36.4 | 70 |
| 15 | 36.4 | 75 |
| 16 | 36.4 | 74 |
| 17 | 36.4 | 69 |
| 18 | 36.4 | 73 |
| 19 | 36.5 | 77 |
| 20 | 36.6 | 58 |
| 21 | 36.6 | 73 |
| 22 | 36.6 | 65 |
| 23 | 36.6 | 74 |
| 24 | 36.6 | 76 |
| 25 | 36.6 | 72 |
| 26 | 36.7 | 78 |
| 27 | 36.7 | 71 |

Bivariate Correlations

Body temp [body_te...
Heart rate [heart_rate]

Variables:

Correlation Coefficients: ☑ Pearson □ Kendall's tau-b □ Spearman

Test of Significance: ● Two-tailed ○ One-tailed

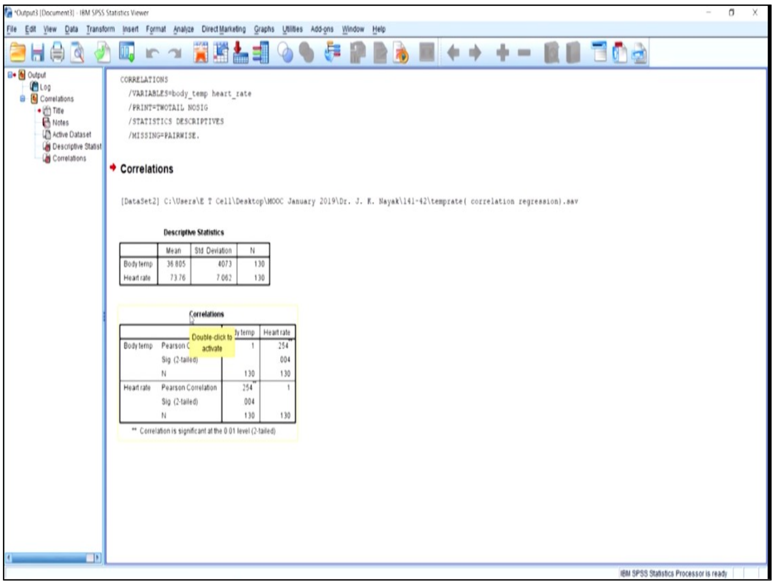
☑ Flag significant correlations

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Visible: 2 of 2 Variables
Options...
Bootstrap...
OK
Paste
Reset
Cancel
Help
Data View
Variable View
Type here to search
11:12 AM
8/21/2018

So I want to take my both the variables hot body temperature and heart rate. In options if you go you can go for means and standard deviations if you want and other things remaining the same. Now let us see is there any relationship between the 2, now body temperature with body temperature the correlation will always be 1 is the same thing because and body temperature and heart rate the correlation is .254. But interestingly the important thing is it is significant at what double ** shows it is significant at 0.01 level.

So you can see your .004 which is much less than .01. So the null hypothesis is rejected. What is the null hypothesis in the case of correlation? You should be able to always write the null hypothesis. The null hypothesis is that there is no correlation or there is no relation between the variables. So this null hypothesis is been rejected. So what is an alternate hypothesis that there is a relationship between the variables. The positive or negative that is separate but there is a relationship.

**(Refer Slide Time: 16:40)**



*Output3 [Document3] - IBM SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Output
Log
Correlations
Title
Notes
Active Dataset
Descriptive Statist
Correlations
Double-click to activate
IBM SPSS Statistics Processor is ready

```
CORRELATIONS
  /VARIABLES=body_temp heart_rate
  /PRINT=TWOTAIL NOSIG
  /STATISTICS DESCRIPTIVES
  /MISSING=PAIRWISE.
```

Correlations

```
[DataSet2] C:\Users\E T Cell\Desktop\MOOC January 2019\Dr. J. K. Nayak\141-42\temprate( correlation regression).sav
```

Descriptive Statistics

| | Mean | Std. Deviation | N |
|---|---|---|---|
| Body temp | 36.805 | .4073 | 130 |
| Heart rate | 73.76 | 7.062 | 130 |

Correlations

| | | dy temp | Heart rate |
|---|---|---|---|
| Body temp | Pearson C | 1 | .254** |
| | Sig. (2-tailed) | | .004 |
| | N | 130 | 130 |
| Heart rate | Pearson Correlation | .254** | 1 |
| | Sig. (2-tailed) | .004 | |
| | N | 130 | 130 |

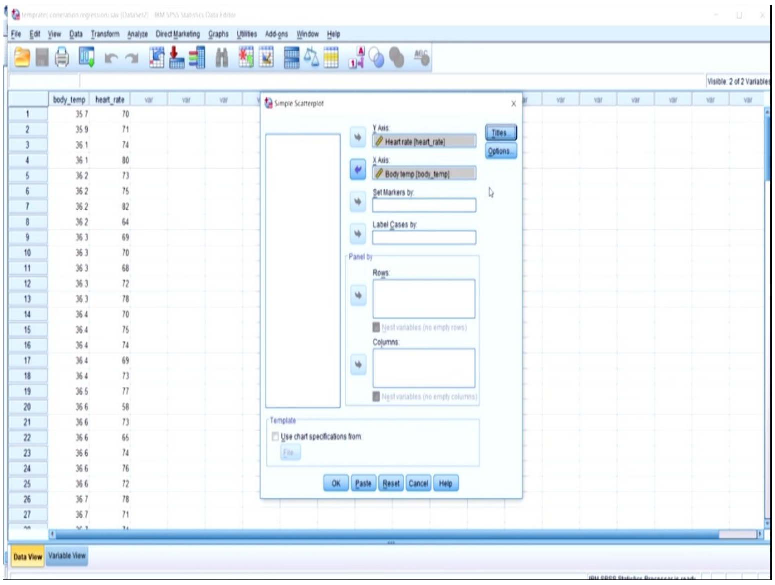
**. Correlation is significant at the 0.01 level (2-tailed).

Now look at the body temperature mean and standard division if you want to use this data you can also use this. Now another example we have this is I think you have understood. Now let us take this one. Now somebody wants to see whether there is a relationship between the amount of distance driven and the number of accidents happening by let us say a bus service provider.

Now let us see what is the relationship coming here so I am taking 2 variables. Number of casualties that is happening by let us say particular bus service organization and the number of billions of kilometre they are driving. Now if you look at the output will go to the output and you see the here the correlation between the 2 is .157. But interestingly if you see the significance value it is .253. So what is happening, this .253 is much above the designated value .05 or .01 which we generally take as the significance level value. So the null hypothesis cannot be rejected. So what is our null hypothesis that there is no relationship between accidents and the amount of kilometre driven?

And yes that is acceptable. That means you cannot say there is any relationship between the 2. You cannot claim that right. So if I want to show you how this would have looked in the scatter diagram because some people wanted. I have explained scatter diagram so I will explain how you show this in scatter diagram, in this relationship. Now go to this Chart Builder or Legacy and go to scatter plot here.

**(Refer Slide Time: 18:29)**



temprate( correlation regression).sav [DataSet2] - IBM SPSS Statistics Data Editor

| | body_temp | heart_rate |
|---|---|---|
| 1 | 35.7 | 70 |
| 2 | 35.9 | 71 |
| 3 | 36.1 | 74 |
| 4 | 36.1 | 80 |
| 5 | 36.2 | 73 |
| 6 | 36.2 | 75 |
| 7 | 36.2 | 82 |
| 8 | 36.2 | 64 |
| 9 | 36.3 | 69 |
| 10 | 36.3 | 70 |
| 11 | 36.3 | 68 |
| 12 | 36.3 | 72 |
| 13 | 36.3 | 78 |
| 14 | 36.4 | 70 |
| 15 | 36.4 | 75 |
| 16 | 36.4 | 74 |
| 17 | 36.4 | 69 |
| 18 | 36.4 | 73 |
| 19 | 36.5 | 77 |
| 20 | 36.6 | 58 |
| 21 | 36.6 | 73 |
| 22 | 36.6 | 65 |
| 23 | 36.6 | 74 |
| 24 | 36.6 | 76 |
| 25 | 36.6 | 72 |
| 26 | 36.7 | 78 |
| 27 | 36.7 | 71 |

Simple Scatterplot

Y Axis: Heart rate [heart_rate]

X Axis: Body temp [body_temp]

Set Markers by:

Label Cases by:

Panel by

Rows:

Nest variables (no empty rows)

Columns:

Nest variables (no empty columns)

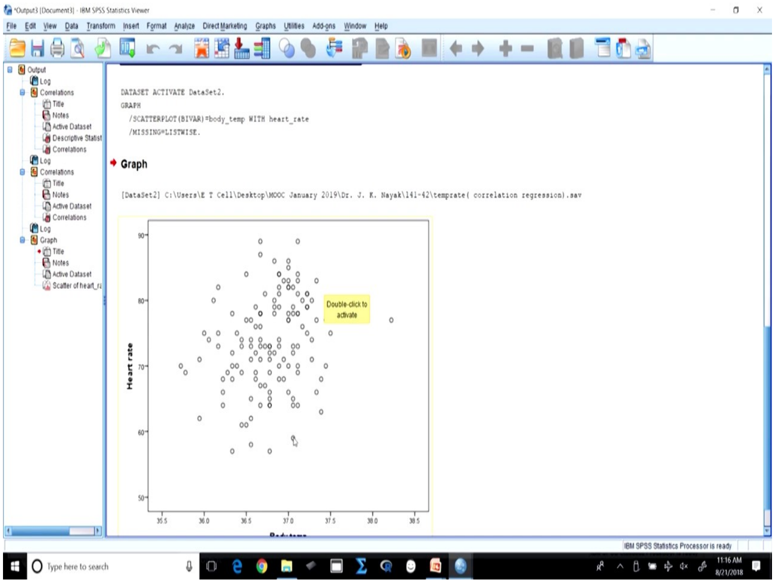
Template

☐ Use chart specifications from:

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Visible: 2 of 2 Variables
Titles...
Options...
File...
OK
Paste
Reset
Cancel
Help
Data View
Variable View

Go to the simple scatter, now take what is your heart rate. Heart rate is my dependent variable so I am taking my heart rate as dependent variable but my body temperature as my X axis my independent variable. If you want to give a title to this you can give it. So suppose some title you want to give you can write here. Now I want to just see it. How does it look.

**(Refer Slide Time: 18:43)**



*Output3 [Document3] - IBM SPSS Statistics Viewer

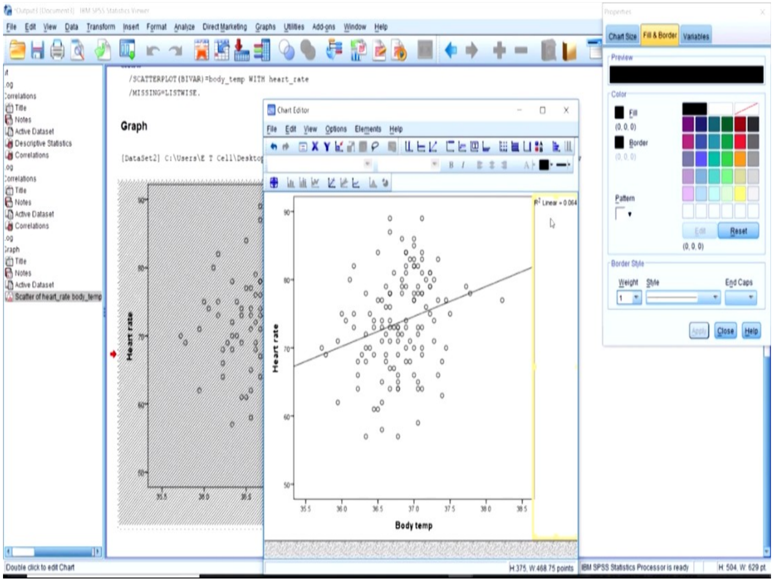
```
DATASET ACTIVATE DataSet2.
GRAPH
  /SCATTERPLOT(BIVAR)=body_temp WITH heart_rate
  /MISSING=LISTWISE.
```

Graph

```
[DataSet2] C:\Users\E T Cell\Desktop\MOOC January 2019\Dr. J. K. Nayak\141-42\temprate( correlation regression).sav
```

So let us go to the output file and if you look at it this is how the variables are scattered the variables whatever data as scattered. So my 2 variables are body temperature x axis, heart rate y axis. And this is my how the data is distributed. If you look at this data and you want to understand then you can you may be able to understand but still to make it more clearer, let us do one thing. Make a double click on this.

**(Refer Slide Time: 19:13)**

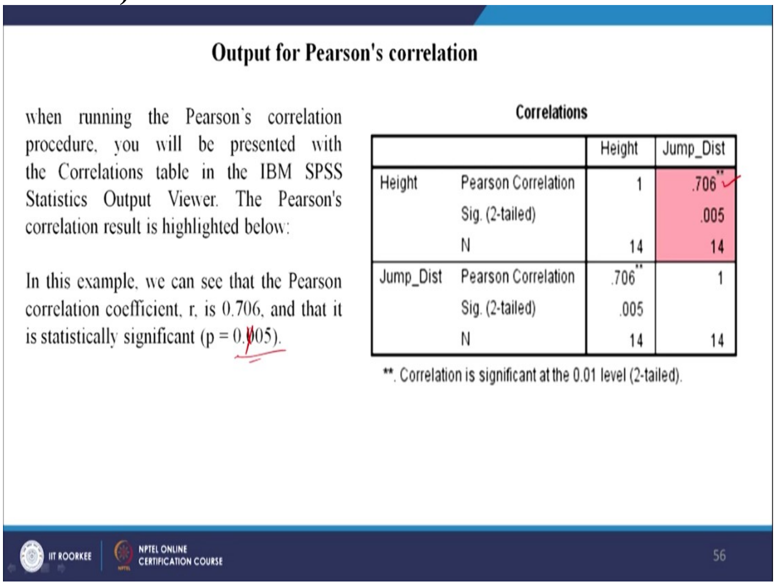


```
/SCATTERPLOT(BIVAR)=body_temp WITH heart_rate
/MISSING=LISTWISE.
```

Now if I make a double click then what I do is I go to this you know fit the total fit. So when I go this I draw a line it draws a line which now helps me to understand the distribution of the data around this line. And if you see here something is written called $r^2$. The r square value is something which I will be explaining you later. It is called the coefficient of determination.

So this $r^2$ is equal to how much 6.4% well. So this tells me well there is some kind of a; if $r^2$ is 6.4. So what is my r? The r you can just say root over of this that means something around. Let us say how much it will be $r^2$ is 6.4 root over of 6.4 is 2.5 or more than that right. So that is you can find out the relationship between r and $r^2$ and thus you say the correlation value is this much. So this is the use of scatter diagram.

**(Refer Slide Time: 20:16)**



## Output for Pearson's correlation

when running the Pearson's correlation procedure, you will be presented with the Correlations table in the IBM SPSS Statistics Output Viewer. The Pearson's correlation result is highlighted below:

In this example, we can see that the Pearson correlation coefficient, r, is 0.706, and that it is statistically significant (p = 0.005).

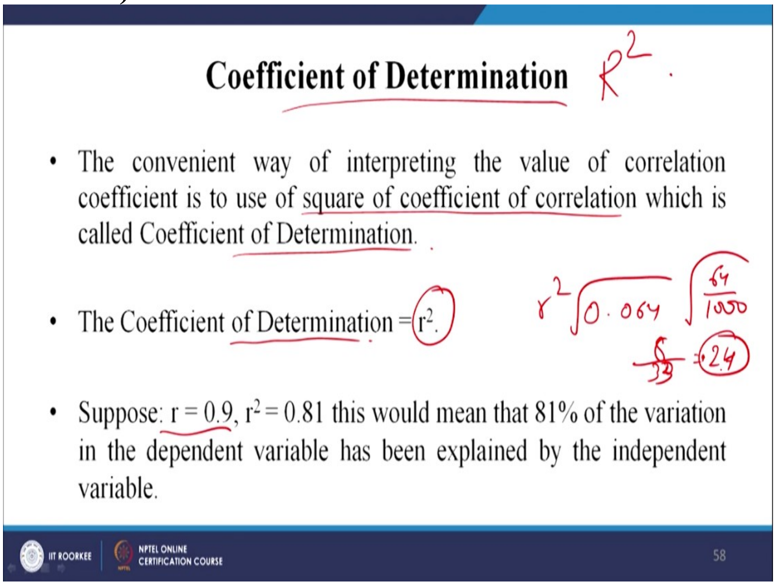
**Correlations**

| | | Height | Jump_Dist |
|---|---|---|---|
| Height | Pearson Correlation | 1 | .706** |
| | Sig. (2-tailed) | | .005 |
| | N | 14 | 14 |
| Jump_Dist | Pearson Correlation | .706** | 1 |
| | Sig. (2-tailed) | .005 | |
| | N | 14 | 14 |

**. Correlation is significant at the 0.01 level (2-tailed).

So it helps you to give it visual look so output for Pearson's Correlation how it writing running in this case you see so height. Supposing this is just a fictitious example height and the jumping distance the Correlation was .706 and it is significant. In this example we can see that the Pearson's correlation coefficient is .706 and it is statistically significant at .05 or taken as .005. But usually it is .05. That means 5%.

**(Refer Slide Time: 20:53)**



## Coefficient of Determination $R^2$.

- The convenient way of interpreting the value of correlation coefficient is to use of square of coefficient of correlation which is called Coefficient of Determination.
- The Coefficient of Determination = $r^2$.
- Suppose: r = 0.9, $r^2$ = 0.81 this would mean that 81% of the variation in the dependent variable has been explained by the independent variable.

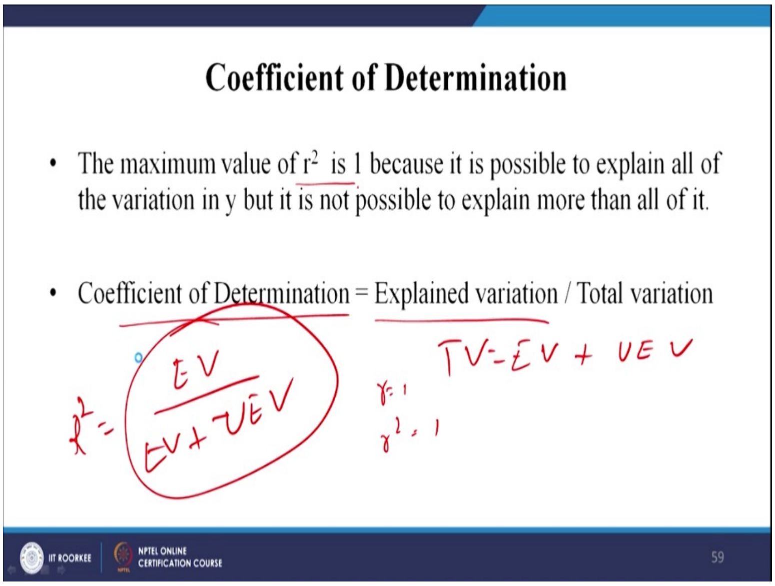
$r^2 \sqrt{0.064}$ $\sqrt{\frac{64}{1000}}$

$\frac{8}{33} = .24$

So that is what it says in our example. What is this coefficient of determination that is what I just said $r^2$. The convenient way of interpreting the value of correlation coefficient is to use the square of coefficient of correlation which is called coefficient of determination. So the coefficient of determination is nothing but this $r^2$. So if r is my 0.9 $r^2$ is 0.81. So similarly if I had a $r^2$ of let us say.064 in my case. So what is my If this is my $r^2$ so what is my r. This is equal to 64 by let us say 1000.

So root over So 8 by let us say how much it 33. So that is somewhere around you are say whatever it comes root over this is this value is equal to 2 point something. So that is what I was saying 2.5 or 2.4 whatever it comes. So this is my r value. So, .24 or .25 whatever it is. Because it cannot be more than – 1 to +1 it has to range in between that.

**(Refer Slide Time: 21:59)**



## Coefficient of Determination

- The maximum value of $r^2$ is 1 because it is possible to explain all of the variation in y but it is not possible to explain more than all of it.
- Coefficient of Determination = Explained variation / Total variation

$$r^2 = \boxed{\frac{EV}{EV+UEV}}$$

$TV = EV + UEV$

$r = 1$

$r^2 = 1$

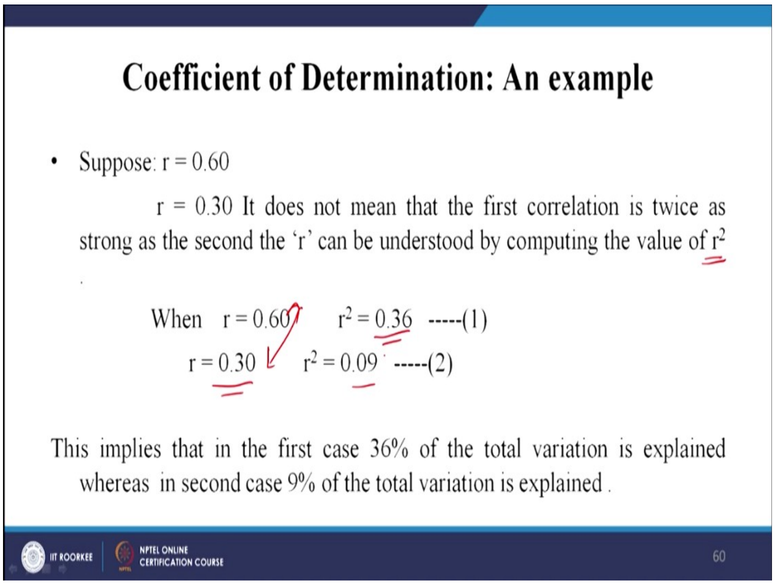
IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 59

Coefficient of determination the maximum value is 1. Because it is possible to explain all of the variation in Y but it is not possible to explain more than all of it. That means what if my r is 1. So r = 1, r square also can be 1 so explained variation coefficient of determination in regression also I will show you. It means it is the ratio of the explained variance by total variance. Now what is this total variance total variance is equal to explained variance plus unexplained variance. So this is my total variance so here I am saying coefficient of determination or

$$r2 = \frac{EV}{EV+UEV}$$

So this is what it seems. Basically theoretically this is the meaning.

**(Refer Slide Time: 22:57)**



## Coefficient of Determination: An example

- Suppose: r = 0.60

r = 0.30 It does not mean that the first correlation is twice as strong as the second the 'r' can be understood by computing the value of $r^2$

When r = 0.60 $r^2 = 0.36$ -----(1)

r = 0.30 $r^2 = 0.09$ -----(2)

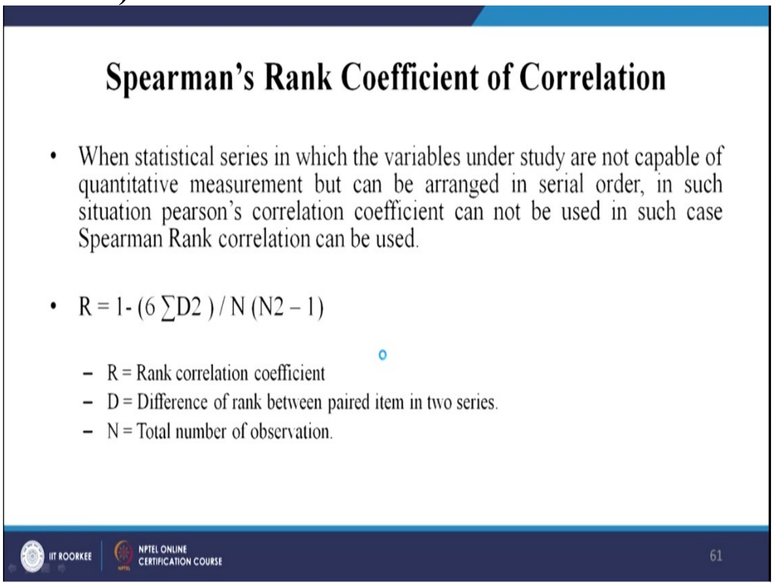
This implies that in the first case 36% of the total variation is explained whereas in second case 9% of the total variation is explained .

IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 60

Suppose r = .6, r = .3 it does not mean that the first correlation is twice as strong as the second. The r can be understood by computing the value of $r^2$. Now let us see r is .6. So what is my $r^2$ is .36. Second case if my r is .30 what is my r square .9. So, here it implies that in the first case 36% of the total variation is explained wherein second case 9% of the total variance is explained which is almost 4 times not 2 times.

From a visual here you could have you would have thought it this 2 times but actually it is 4 times the explanation that it is giving you is 4 times. So well what I will do is there is another method. This is for the coefficient of correlation. You have understood.

**(Refer Slide Time: 23:57)**



## Spearman's Rank Coefficient of Correlation

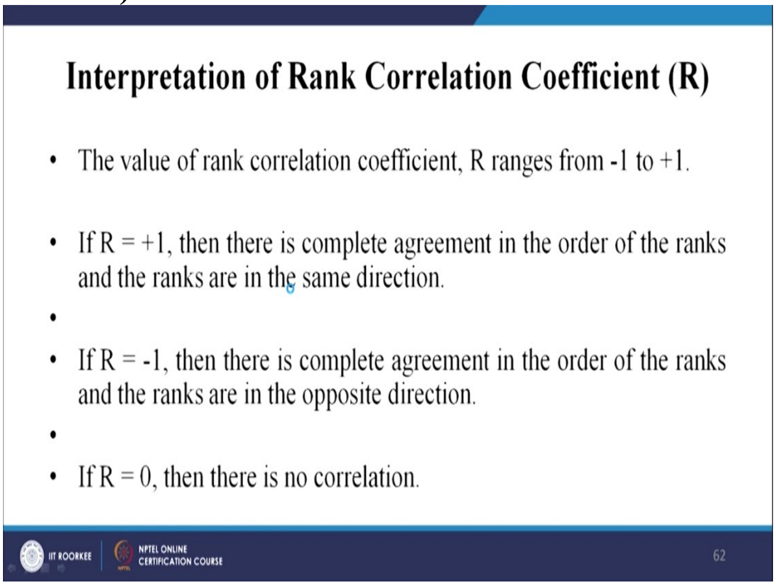
- When statistical series in which the variables under study are not capable of quantitative measurement but can be arranged in serial order, in such situation pearson's correlation coefficient can not be used in such case Spearman Rank correlation can be used.
- $R = 1- (6 \sum D2 ) / N (N2 - 1)$
  - R = Rank correlation coefficient
  - D = Difference of rank between paired item in two series.
  - N = Total number of observation.

Thus another technique which is also used for a non-parametric data and this test is called a spearman's correlation which is called Spearman's Rank coefficient of correlation. I have already done it earlier also. But for the benefit of many of the viewers who have not followed that I can just brief you. This Spearman Rank correlation coefficient is nothing but a study in which the data is in a non-parametric nature. So in which the variables are not capable of a quantitative measurement but can be arranged in a serial order.

So they do not follow a normal distribution because they are not in a continuous manner. They are not collected in a continuous manner. So that is one of the assumptions that it should be in ratio an interval scale for a correlation. But here it is not in ratio or an interval scale rather it is in the order that is in a ordinal scale. In such situation the Pearson correlation coefficient cannot be used. And we use this Spearman correlation. So this is the formula

$R = 1-(6\sum D^2)/N(N^2-1)$

**(Refer Slide Time: 24:59)**



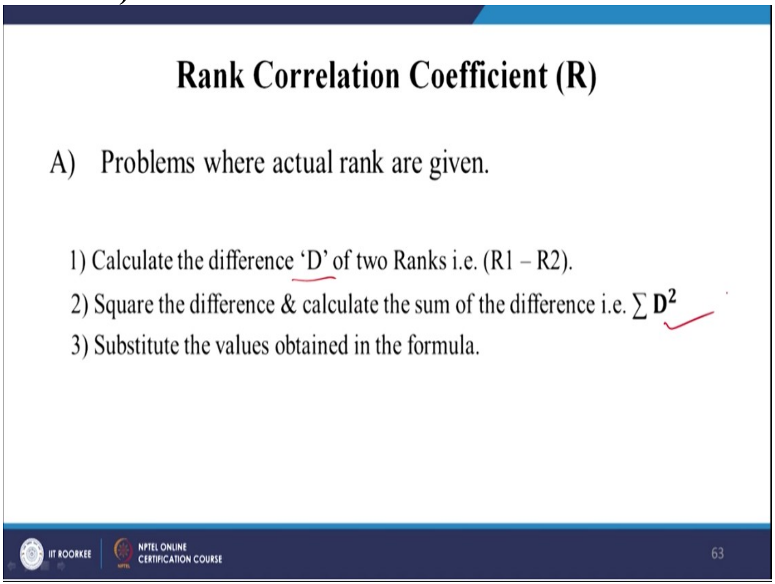
## Interpretation of Rank Correlation Coefficient (R)

- The value of rank correlation coefficient, R ranges from -1 to +1.
- If R = +1, then there is complete agreement in the order of the ranks and the ranks are in the same direction.
- 
- If R = -1, then there is complete agreement in the order of the ranks and the ranks are in the opposite direction.
- 
- If R = 0, then there is no correlation.

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 62

So, this other things remaining more or less the same.

**(Refer Slide Time: 25:03)**



## Rank Correlation Coefficient (R)

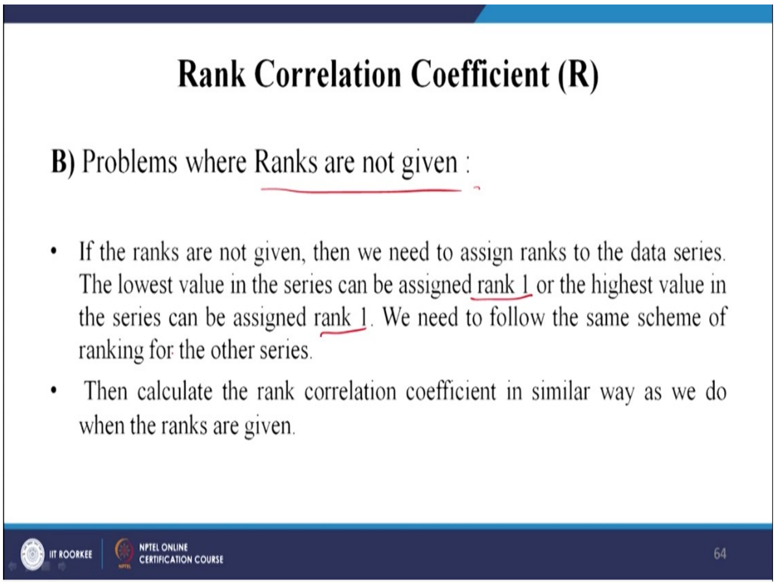
A) Problems where actual rank are given.

1) Calculate the difference 'D' of two Ranks i.e. (R1 – R2).
2) Square the difference & calculate the sum of the difference i.e. $\sum \mathbf{D^2}$
3) Substitute the values obtained in the formula.

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 63

So let us see how do you do it. Now calculate the difference D of 2 ranks R1 - R2 and square the difference and calculate that D square.

**(Refer Slide Time: 25:15)**



## Rank Correlation Coefficient (R)

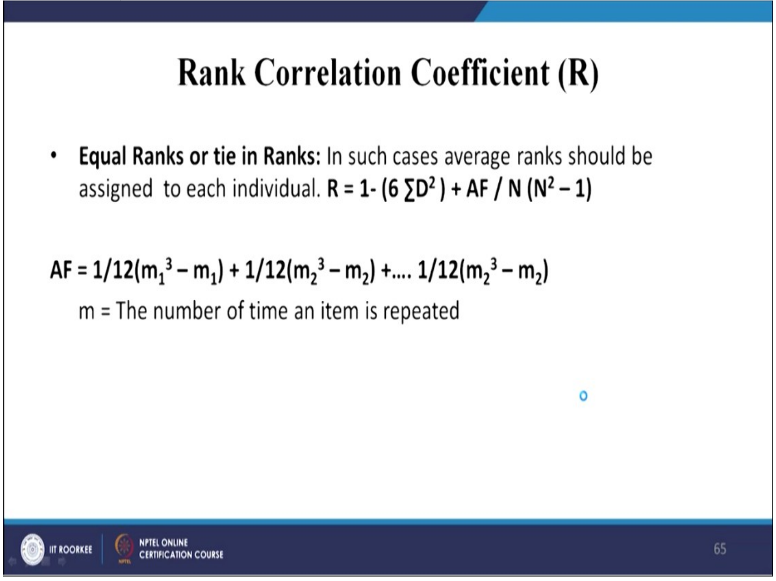
**B)** Problems where Ranks are not given :

- If the ranks are not given, then we need to assign ranks to the data series. The lowest value in the series can be assigned rank 1 or the highest value in the series can be assigned rank 1. We need to follow the same scheme of ranking for the other series.
- Then calculate the rank correlation coefficient in similar way as we do when the ranks are given.

Substitute the values obtained in the formula that was shown to you. Problems where ranks are not given to you, if the ranks are not given then we need to assign the ranks if ranks are not given just assign ranks. So, by maybe the number the score of somebody else you can assign ranks. The lowest value will be ranked 1 or the highest value can be ranked 1 whatever order you want to take.

We need to follow the same scheme of ranking for the other series also. Suppose there are 2 variables 1 and 2. So if you are using in ascending order for 1 please use ascending order for the other one also. If you are using descending use descending for both.

**(Refer Slide Time: 25:55)**



## Rank Correlation Coefficient (R)

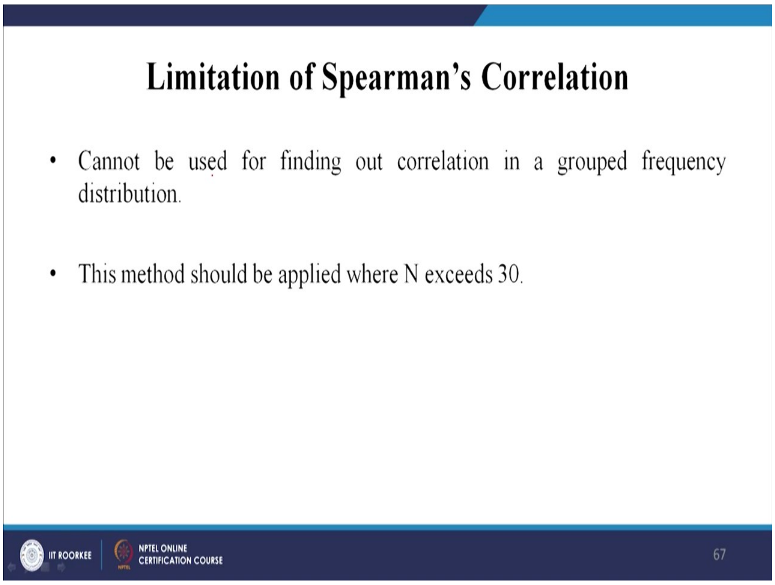
- **Equal Ranks or tie in Ranks:** In such cases average ranks should be assigned to each individual. **$R = 1- (6 \sum D^2 ) + AF / N (N^2 – 1)$**

**$AF = 1/12(m_1^3 – m_1) + 1/12(m_2^3 – m_2) +.... 1/12(m_2^3 – m_2)$**

m = The number of time an item is repeated

Then calculate the rank correlation coefficient. So this is the formula.

**(Refer Slide Time: 26:01)**



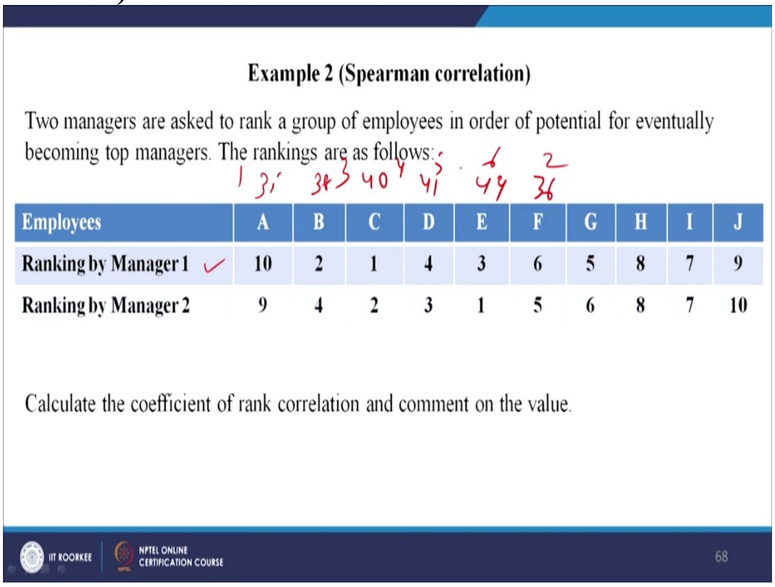
## Limitation of Spearman's Correlation

- Cannot be used for finding out correlation in a grouped frequency distribution.
- This method should be applied where N exceeds 30.

IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE
67

Let me take it. If I have what the limitation of this, it cannot be used for finding out correlation in a group frequency distribution and this method should be applied where N exceeds 30.

**(Refer Slide Time: 26:11)**



### Example 2 (Spearman correlation)

Two managers are asked to rank a group of employees in order of potential for eventually becoming top managers. The rankings are as follows:

1 35 38 3 40 4 41 5 44 6 36 2

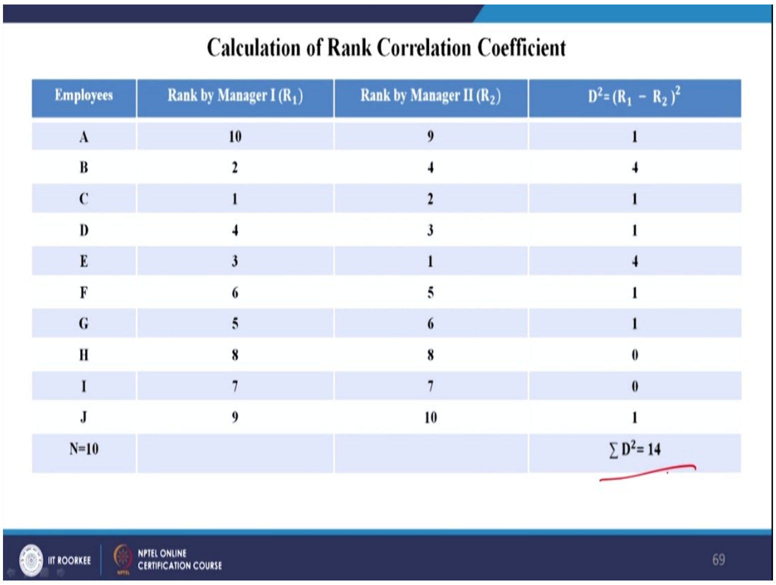
| Employees | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| Ranking by Manager 1 ✓ | 10 | 2 | 1 | 4 | 3 | 6 | 5 | 8 | 7 | 9 |
| Ranking by Manager 2 | 9 | 4 | 2 | 3 | 1 | 5 | 6 | 8 | 7 | 10 |

Calculate the coefficient of rank correlation and comment on the value.

IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE
68

Now this is the case ranking by manager one of several employees A to J, there are 10 employees. But Manager1 gives these ranks to these people. Manager2 gives this rank. Suppose it would not have given rank to you and it would have given you some score let us say 35 38 40 41 44 so 36, so what would have done is you would have made them a rank how. Now for example I am starting with the lowest listed 1. Let us say 2. Let us say 3 , 4, 5, 6. So this is how you do and Manager II also has given some ranks.

**(Refer Slide Time: 26:50)**



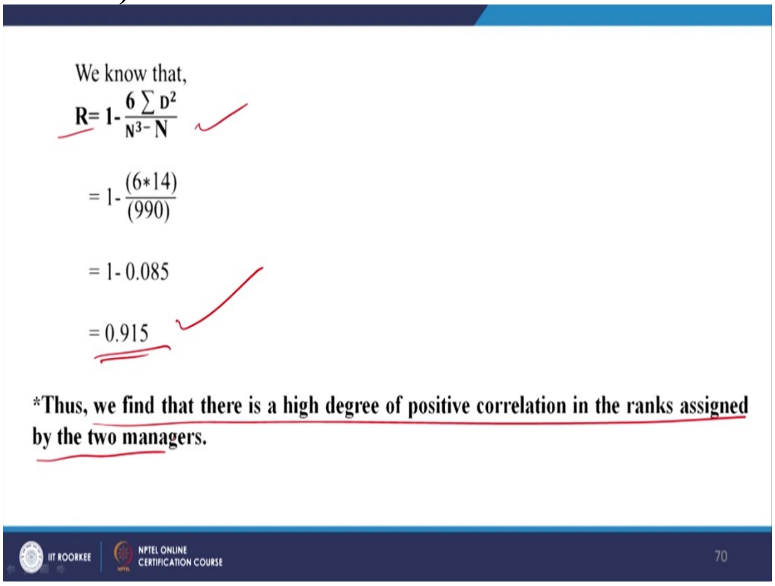
Calculation of Rank Correlation Coefficient

| Employees | Rank by Manager I ($R_1$) | Rank by Manager II ($R_2$) | $D^2 = (R_1 - R_2)^2$ |
|---|---|---|---|
| A | 10 | 9 | 1 |
| B | 2 | 4 | 4 |
| C | 1 | 2 | 1 |
| D | 4 | 3 | 1 |
| E | 3 | 1 | 4 |
| F | 6 | 5 | 1 |
| G | 5 | 6 | 1 |
| H | 8 | 8 | 0 |
| I | 7 | 7 | 0 |
| J | 9 | 10 | 1 |
| N=10 | | | $\sum D^2 = 14$ |

IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE
69

Now this is the rank by Manager I and this is rank by Manager II. I am finding the D the difference between the 2 the difference between the first 1 is 1. So, I will take $1^2$, 2-4=- -$2^2$= - 1 -2=1 it goes on. So my $D^2$ is 14.

**(Refer Slide Time: 27:12)**



We know that,

$$R = 1-\frac{6\sum D^2}{N^3-N}$$

$$= 1-\frac{(6*14)}{(990)}$$

$$= 1-0.085$$

$$= 0.915$$

*Thus, we find that there is a high degree of positive correlation in the ranks assigned by the two managers.

IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE
70

If I have my $D^2$ we know that R with the formula for the Spearman correlation

$$R = 1-(6\sum D^2)/N^3-N).$$

This is hardly it makes a difference it actually does not make much of a difference at all right. So if I do this is coming .915. So, we say does we find that there is a high degree of positive correlation in the ranks assigned by the 2 managers why it is helpful because it is helpful that it says that 2 managers are thinking alike.

So, if you are thinking alike otherwise we would have said that managers are thinking differently. So, this is about the Spearman's correlation. So earlier we discussed about the

Pearson's Correlation. We saw how to do that Pearson's Correlation. And we also saw how today how to do the Spearman’s correlation for a non metric variable or non metric data. So this is all for the day. Thank you so much.