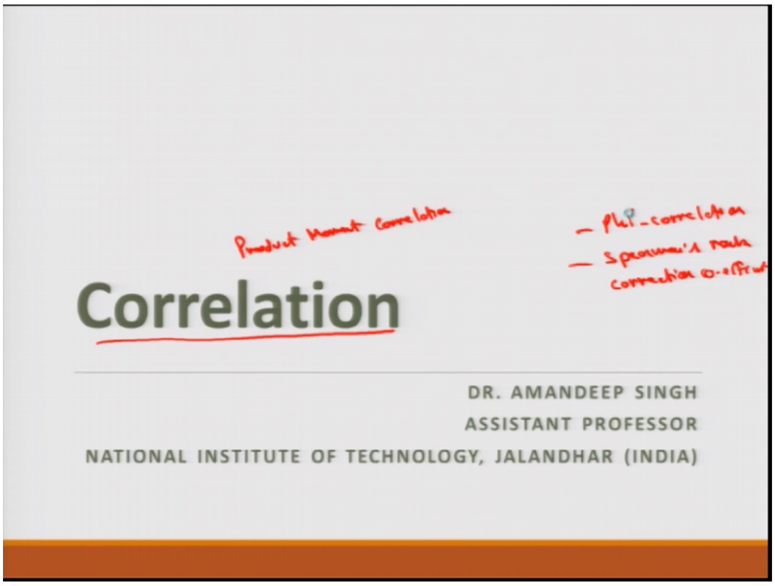
Practitioners Course in Descriptive, Predictive and Prescriptive Analytics

Prof. Deepu Philip
Dr. Amandeep Singh Oberoi
Mr. Sanjeev Newar
Department of Industrial & Management Engineering
Indian Institute of Technology, Kanpur
National Institute of Technology, Jalandhar

Lecture - 20
Correlation

So, welcome back to the course on Analytics.

(Refer Slide Time: 00:17)



Product Moment Correlation

– Phi-correlation

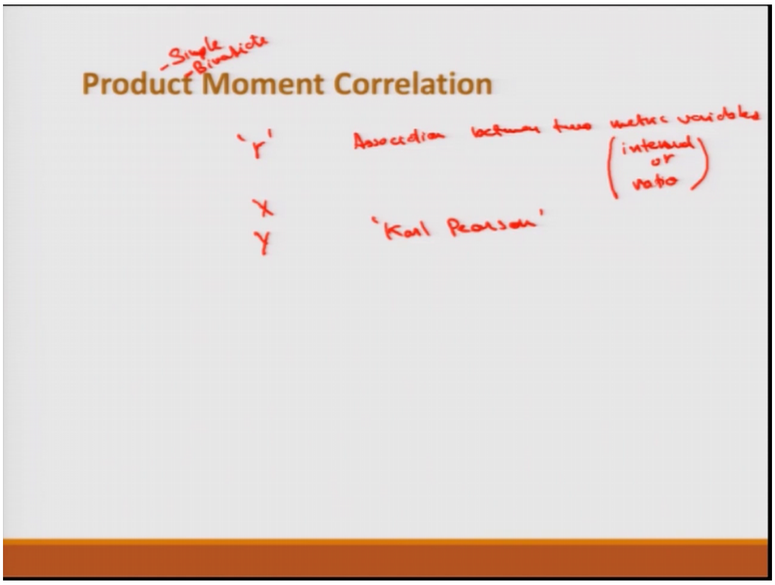
– Spearman's rank correlation co-efficient

Correlation

DR. AMANDEEP SINGH
ASSISTANT PROFESSOR
NATIONAL INSTITUTE OF TECHNOLOGY, JALANDHAR (INDIA)

In this lecture, I will like to discuss the correlations; correlation specifically product moment correlation. Further phi correlation and Spearman's the rank correlation coefficient would be discussed this is a quick session to just have an overview of product moment correlation.

(Refer Slide Time: 00:57)



-Simple
-Bivariate

Product Moment Correlation

'r' Association between two metric variables
(interval or ratio)

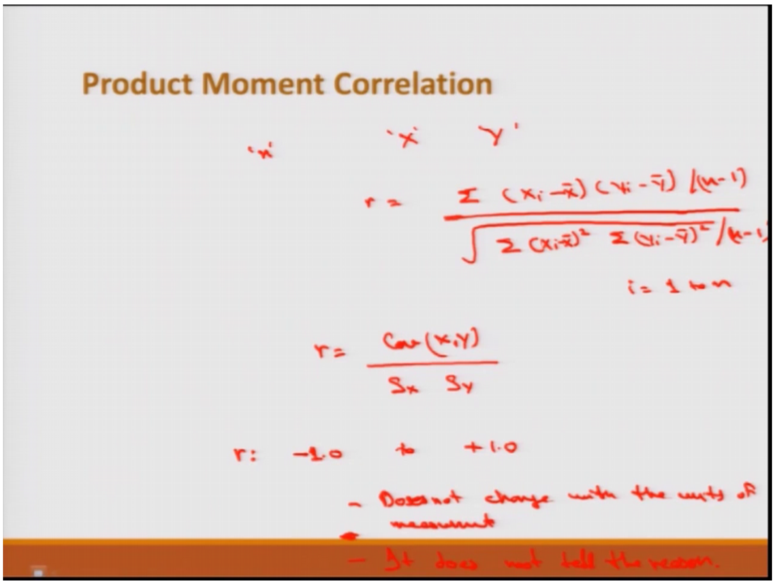
X
Y

'Karl Pearson'

Product moment correlation that is small r this is noted by small r that summarises the strength of association between two metric variables; association between two metric variables. Mmetric means it could be either interval or ratio scale; so say the variables are X and Y.

So, it is an index that is used to determine whether a linear or straight line relationship exist between X and Y and it was originally proposed by Karl Pearson. So, this coefficient r is also known as Karl Pearson correlation coefficient; so, it is also referred as simple correlation; simple correlation or bivariate correlation because two variables are related here.

(Refer Slide Time: 02:32)



Product Moment Correlation

'n' 'x' 'y'

$$r = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y}) / (n-1)}{\sqrt{\Sigma (x_i - \bar{x})^2 \; \Sigma (y_i - \bar{y})^2 / (n-1)}}$$

$i = 1$ to n

$$r = \frac{Cov(x, y)}{S_x \; S_y}$$

r: −1.0 to +1.0

- Does not change with the units of measurement
- It does not tell the reason.

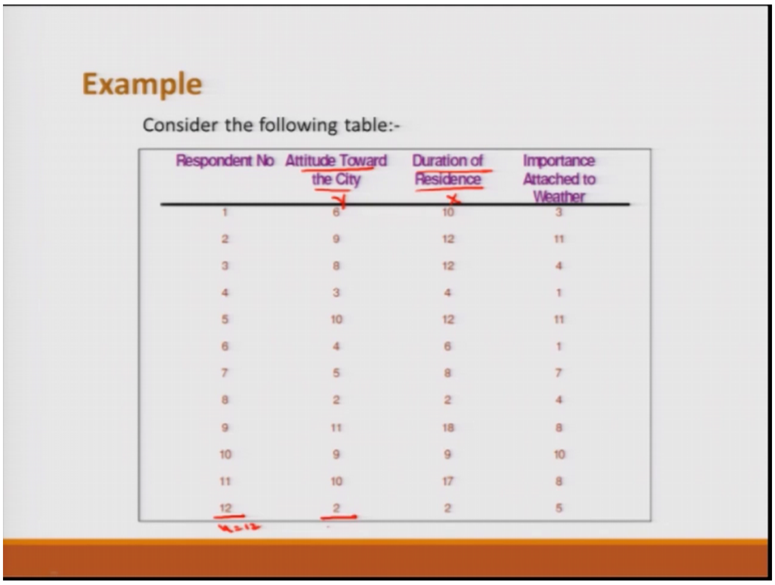
So, from a sample of n observations of x and y the product moment correlation or Karl Pearson correlation can be calculated using this relation; r is equal to summation the difference of X i to X bar and the difference of Y i to Y bar over sum of squares for X i and sum of squares for Y i.

So, this all i r from 1 to n; so, if we see this if we divide both of these by degrees of freedom by n minus 1; divide this by n minus 1, we can see that the relation add in the numerator is covariance of XY and the relation the numerator is the standard deviation for X and standard deviation for Y if I take this into the under root.

So, r is nothing, but covariance of XY over standard deviation for X and standard deviation for Y; r varies between minus 1 to plus 1 plus 1 be between these perfect positive correlation minus 1 means perfect negative correlation. The correlation coefficient between two variables will be the same regardless of their underlying units of measurement correlation does not change; correlation does not change with the units of measurement.

Also it does not tell the reason it does not tell the reason that why there is positive correlation or negative correlation; it just tells that yes it is varying positively.

(Refer Slide Time: 05:21)



Example

Consider the following table:-

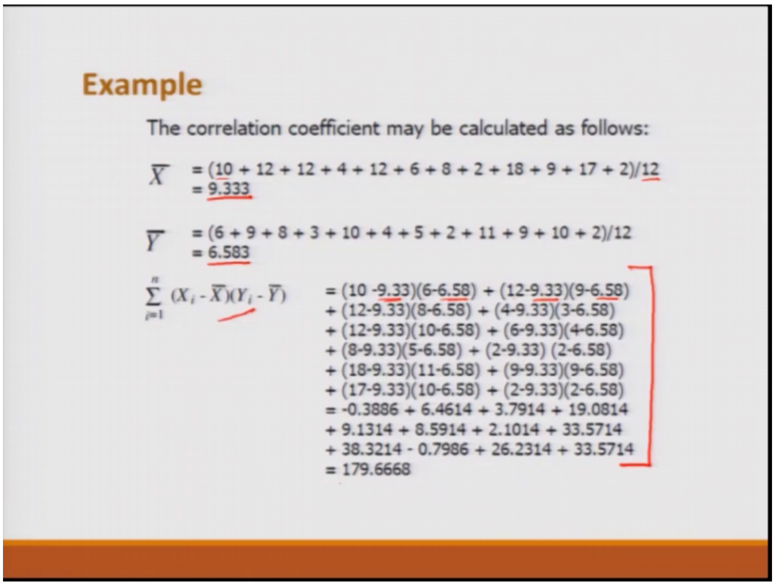
Respondent No	Attitude Toward the City Y	Duration of Residence X	Importance Attached to Weather
1	6	10	3
2	9	12	11
3	8	12	4
4	3	4	1
5	10	12	11
6	4	6	1
7	5	8	7
8	2	2	4
9	11	18	8
10	9	9	10
11	10	17	8
12	2	2	5

n=12

Let me take an example for this. So, we have the respondent numbers here they are 12 respondents attitude towards city is scaled here. And duration of residence in the city is given importance attached to weather is given there are three variables and n is equal to 12. So, what do we do?

So, let us try to find the correlation between the attitude towards the city and duration of residence. Let me put this as Y variable and this as X variable; so, first we will calculate X bar and Y bar here Y bar comes down to that is calculated here.

(Refer Slide Time: 06:03)



Example

The correlation coefficient may be calculated as follows:

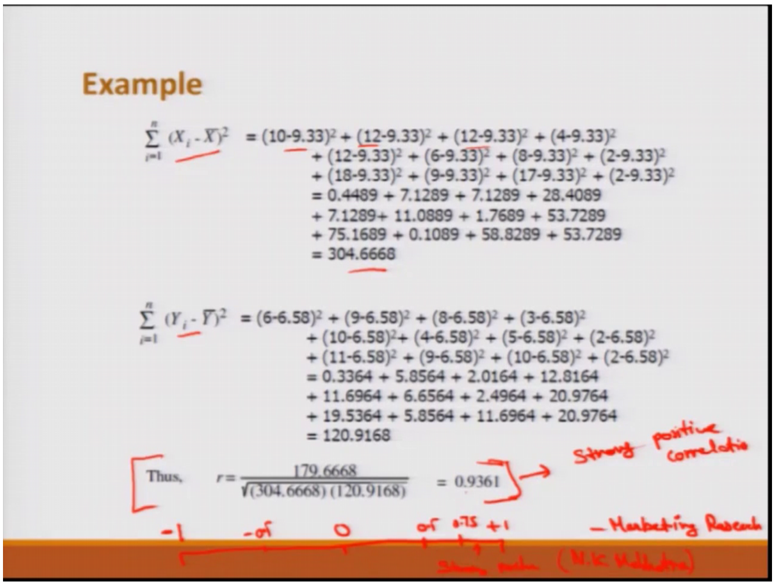
$$\overline{X} = (10 + 12 + 12 + 4 + 12 + 6 + 8 + 2 + 18 + 9 + 17 + 2)/12$$
$$= 9.333$$

$$\overline{Y} = (6 + 9 + 8 + 3 + 10 + 4 + 5 + 2 + 11 + 9 + 10 + 2)/12$$
$$= 6.583$$

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) &= (10 - 9.33)(6-6.58) + (12-9.33)(9-6.58) \\ &+ (12-9.33)(8-6.58) + (4-9.33)(3-6.58) \\ &+ (12-9.33)(10-6.58) + (6-9.33)(4-6.58) \\ &+ (8-9.33)(5-6.58) + (2-9.33)\ (2-6.58) \\ &+ (18-9.33)(11-6.58) + (9-9.33)(9-6.58) \\ &+ (17-9.33)(10-6.58) + (2-9.33)(2-6.58) \\ &= -0.3886 + 6.4614 + 3.7914 + 19.0814 \\ &+ 9.1314 + 8.5914 + 2.1014 + 33.5714 \\ &+ 38.3214 - 0.7986 + 26.2314 + 33.5714 \\ &= 179.6668\end{aligned}$$

X bar is average of all the values here, Y bar is average of all the values here. So, this X bar is sum divided by 12; so, this comes down to 9.33, Y bar is 6.853. We can calculate X i minus X bar into Y i minus Y bar in this way 10 minus X bar into 6 minus Y bar 12 minus X bar into 9 minus Y bar. So, this whole calculation is done here and this value comes down to 179.

(Refer Slide Time: 06:41)



Example

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\bar{X})^2 &= (10-9.33)^2+(12-9.33)^2+(12-9.33)^2+(4-9.33)^2\\ &+(12-9.33)^2+(6-9.33)^2+(8-9.33)^2+(2-9.33)^2\\ &+(18-9.33)^2+(9-9.33)^2+(17-9.33)^2+(2-9.33)^2\\ &=0.4489+7.1289+7.1289+28.4089\\ &+7.1289+11.0889+1.7689+53.7289\\ &+75.1689+0.1089+58.8289+53.7289\\ &=304.6668\end{aligned}$$

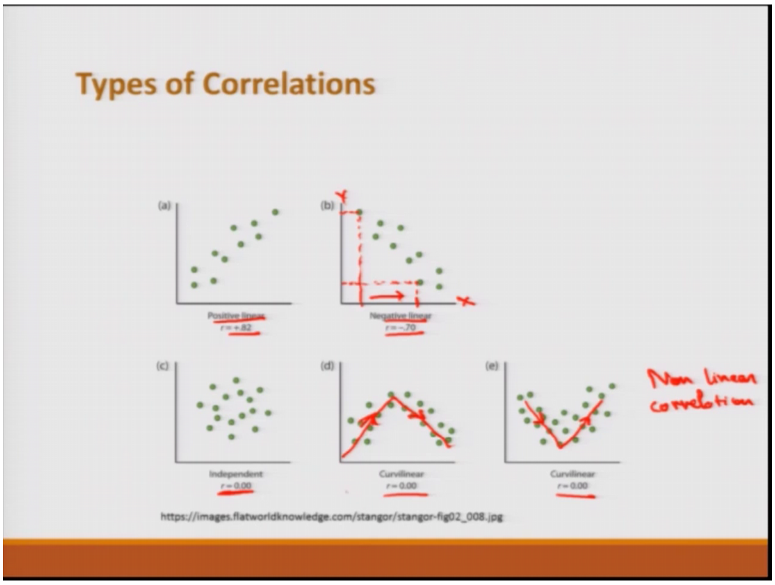
$$\begin{aligned}\sum_{i=1}^{n}(Y_i-\bar{Y})^2 &= (6-6.58)^2+(9-6.58)^2+(8-6.58)^2+(3-6.58)^2\\ &+(10-6.58)^2+(4-6.58)^2+(5-6.58)^2+(2-6.58)^2\\ &+(11-6.58)^2+(9-6.58)^2+(10-6.58)^2+(2-6.58)^2\\ &=0.3364+5.8564+2.0164+12.8164\\ &+11.6964+6.6564+2.4964+20.9764\\ &+19.5364+5.8564+11.6964+20.9764\\ &=120.9168\end{aligned}$$

Thus, $r=\dfrac{179.6668}{\sqrt{(304.6668)(120.9168)}} = 0.9361$ → Strong positive correlatio

Similarly, we calculate X i minus X bar square for all the observations Y i minus Y bar square and we calculate this r. So, these calculations are taken from the book marketing research by NK Malhotra you can refer the book for more details. So, this is very sample correlation.

So, what is the value? The value is coming 0.93 which means the correlation is very strong if I divide the correlation from minus 1 to plus 1 0 here in between if I divide this correlation 0 means no correlation minus 1 means perfect negative correlation plus 1 is perfect positive correlation. And if I divide this further 0.5 minus 0.5; if this is more than 0.75 this say that this is strong positive.

(Refer Slide Time: 08:06)



Types of Correlations

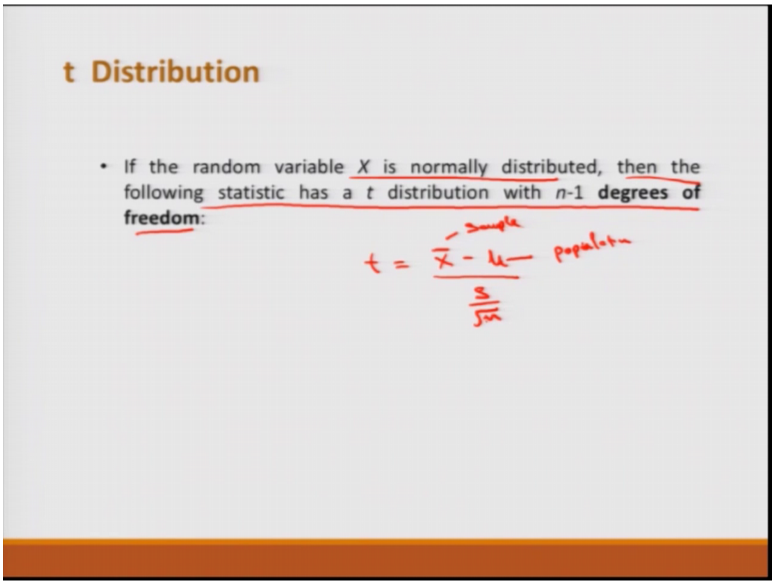
https://images.flatworldknowledge.com/stangor/stangor-fig02_008.jpg

So, this exhibits strong positive correlation types correlations this is positive correlation and linear strong positive correlation, this is negative correlation you can see that by increasing value of X here the Y is decreasing at this value of X Y is large at the larger value of X Y is small; so, this is a negative correlation.

So, in the previous example we can say the correlation is strong; that means, we can say that the attitude towards city is very highly correlated with the duration of residence. The length of residence or the duration of resident the longer the resident had stayed in the city the positive he has the attitude towards the city. So, he is quite loyal or quite loving toward his city.

So, this is no correlation r 0 this is again r 0; there is increasing and decreasing correlation these are Karl non-linear correlation. So, this is the drawback or pitfall of the correlation there is a relation increasing then decreasing, but the correlation coefficient says I do not know if there exist some relation or not. So, this is not explained by the correlation coefficient. So, for this we need to fit some regression model here.

(Refer Slide Time: 09:53)



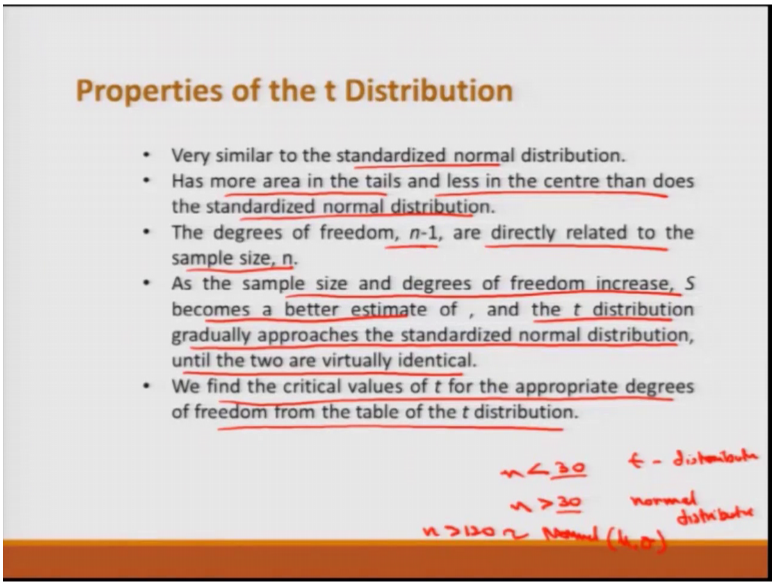
t Distribution

- If the random variable X is normally distributed, then the following statistic has a t distribution with n-1 **degrees of freedom**:

$$t = \frac{\overset{\text{sample}}{\bar{X}} - \mu \text{ — population}}{\frac{s}{\sqrt{n}}}$$

So, before talking more about correlation I will have to discuss the t distribution if random variable X is normally distributed then the following statistic has t distribution with n minus 1 degrees of freedom.

(Refer Slide Time: 10:48)



Properties of the t Distribution

- Very similar to the standardized normal distribution.
- Has more area in the tails and less in the centre than does the standardized normal distribution.
- The degrees of freedom, n-1, are directly related to the sample size, n.
- As the sample size and degrees of freedom increase, S becomes a better estimate of , and the t distribution gradually approaches the standardized normal distribution, until the two are virtually identical.
- We find the critical values of t for the appropriate degrees of freedom from the table of the t distribution.

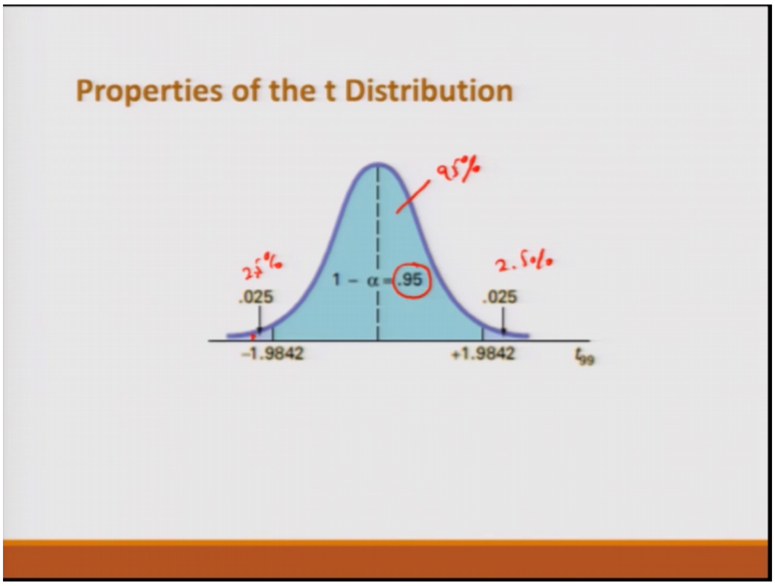
$n < \underline{30}$ t - distribution

$n > \underline{30}$ normal distribution

$n > 120 \sim$ Normal (μ, σ)

Where t is calculated as X bar that is for sample minus population mean by standard deviation for the sample. So, t is calculated as X bar that is for sample minus mu that is for a population by standard deviation because it is for sample we have divided this by under root n. So, very similar to the standardised normal distribution the t distribution behaves it has more area in tails and less in the centre than the standardized normal distribution, we will see the t distribution here.

(Refer Slide Time: 11:00)

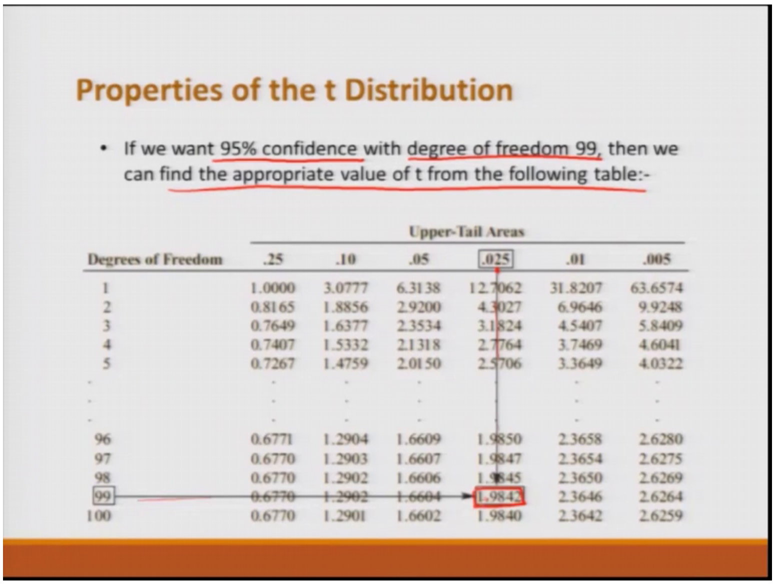


Properties of the t Distribution

This is t distribution. So, these are the value at minus 1.98 deviate value; we have 0.25 area on the left hand side and 0.25 percent of area on the right hand side. So, the confidence interval here is 95 percent; the degrees of freedom n minus 1 are directly related to the sample size n. The this is freedom minus 1 is here because we know one statistic one of the parameter is known as the sample size and degrees of freedom increase S becomes a better estimate and the t distribution gradually approach is the standardised normal distribution for large sample size until that two are virtually identical.

We find that the critical values of t for the appropriate degrees of freedom from the table of t distribution. So, if sample size is generally less than 30; we applied t distribution if it is greater than 30, it is normal distribution. So, this approach is normal distribution; in some cases some researchers say that this value should be 120 and greater than 120 is exactly normal distribution.

(Refer Slide Time: 12:42)



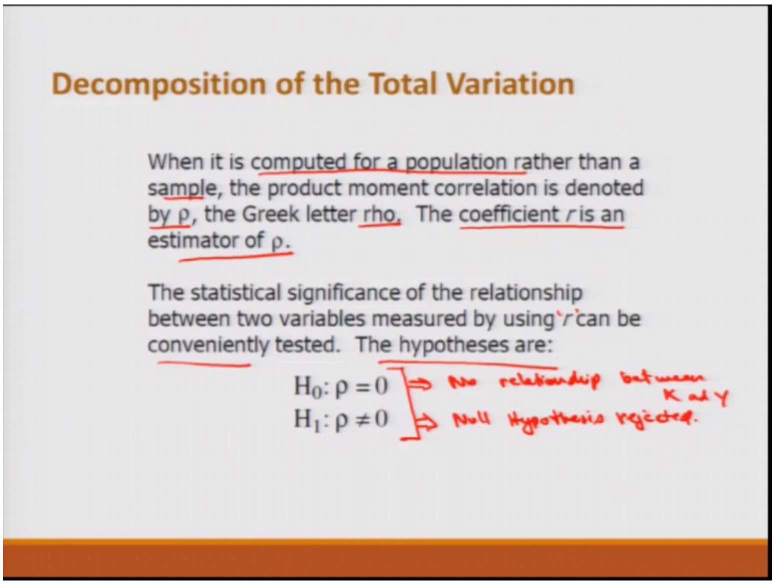
Properties of the t Distribution

- If we want 95% confidence with degree of freedom 99, then we can find the appropriate value of t from the following table:-

	Upper-Tail Areas					
Degrees of Freedom	.25	.10	.05	.025	.01	.005
1	1.0000	3.0777	6.3138	12.7062	31.8207	63.6574
2	0.8165	1.8856	2.9200	4.3027	6.9646	9.9248
3	0.7649	1.6377	2.3534	3.1824	4.5407	5.8409
4	0.7407	1.5332	2.1318	2.7764	3.7469	4.6041
5	0.7267	1.4759	2.0150	2.5706	3.3649	4.0322
.	.	.	.		.	.
.	.	.	.		.	.
.	.	.	.		.	.
96	0.6771	1.2904	1.6609	1.9850	2.3658	2.6280
97	0.6770	1.2903	1.6607	1.9847	2.3654	2.6275
98	0.6770	1.2902	1.6606	1.9845	2.3650	2.6269
99	0.6770	1.2902	1.6604	1.9842	2.3646	2.6264
100	0.6770	1.2901	1.6602	1.9840	2.3642	2.6259

So, if you want 95 percent confidence interval with degrees of freedom 99 then we can find the appropriate value of t from the following table. So, at 95 percent confidence interval 95 percent confidence interval as we just saw at 95 percent confidence interval with two tail test here, the area is distributed 2.5 percent on this side 2.5 percent on this side and 95 percent is here. So, at 95 percent for two tail test this is the value; so, degrees of freedom here 99 this is the value of t statistic here.

(Refer Slide Time: 13:34)



Decomposition of the Total Variation

When it is computed for a population rather than a sample, the product moment correlation is denoted by ρ, the Greek letter rho. The coefficient *r* is an estimator of ρ.

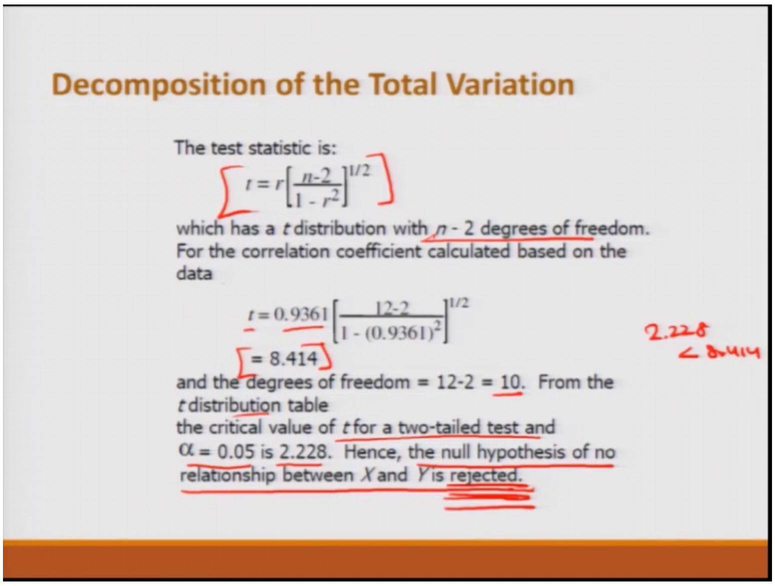
The statistical significance of the relationship between two variables measured by using '*r*' can be conveniently tested. The hypotheses are:

$$H_0: \rho = 0 \Rightarrow \text{No relationship between X and Y}$$

$$H_1: \rho \neq 0 \Rightarrow \text{Null Hypothesis rejected.}$$

So, decomposition of total variation when the product moment correlation is computed for a population rather than sample it is denoted by the letter rho the Greek letter here. The coefficient r is an estimator of rho then the statistical significance of the relationship between two variables measured by using r can be conveniently tested. So, that is make the hypothesis here H naught which is null hypothesis that implies no relationship between X and Y. So, this is that null hypothesis rejected that is no relationship hypothesis is rejected.

(Refer Slide Time: 14:39)



Decomposition of the Total Variation

The test statistic is:

$$t = r\left[\frac{n-2}{1-r^2}\right]^{1/2}$$

which has a t distribution with n - 2 degrees of freedom. For the correlation coefficient calculated based on the data

$$t = 0.9361\left[\frac{12-2}{1-(0.9361)^2}\right]^{1/2}$$

$$= 8.414$$

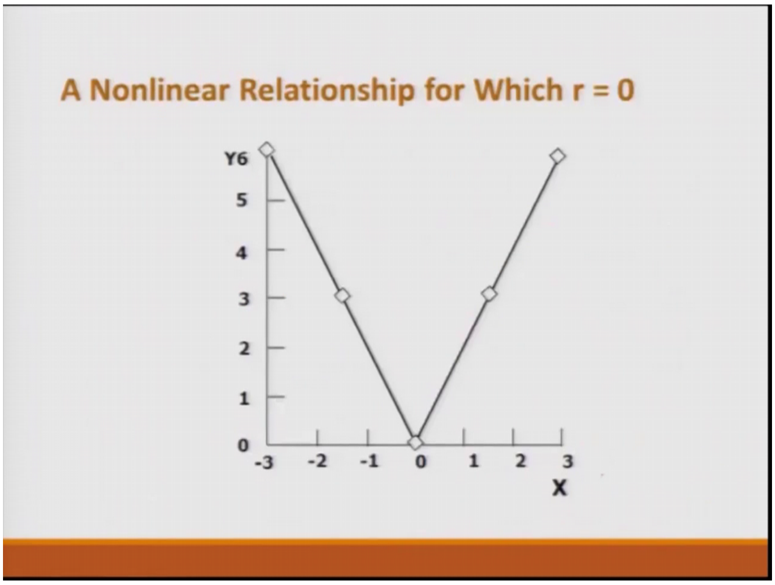
and the degrees of freedom = 12-2 = 10. From the t distribution table the critical value of t for a two-tailed test and $\alpha = 0.05$ is 2.228. Hence, the null hypothesis of no relationship between X and Y is rejected.

$2.228 < 8.414$

So, the test statistic is evaluated using this relation t is equal to r into n minus 2 by 1 minus r square under root which has t distribution with n minus 2 degrees of freedom. So, we calculate the value of t here 0.93 was the correlation coefficient value which we calculated here 0.9361 and the other values we put here 12 minus 2 so, on.

So, the t statistic value is 8.414 and for the 10 degrees of freedom from the t distribution table the critical value of the two tailed test for 95 percent confidence interval is 2.228 hence the critical value is way less than 2.228 is less than 8.414. Hence the null hypothesis of no relationship between X and Y is rejected; that means, there existent relationship. So, this is one of the application of the t distribution here.

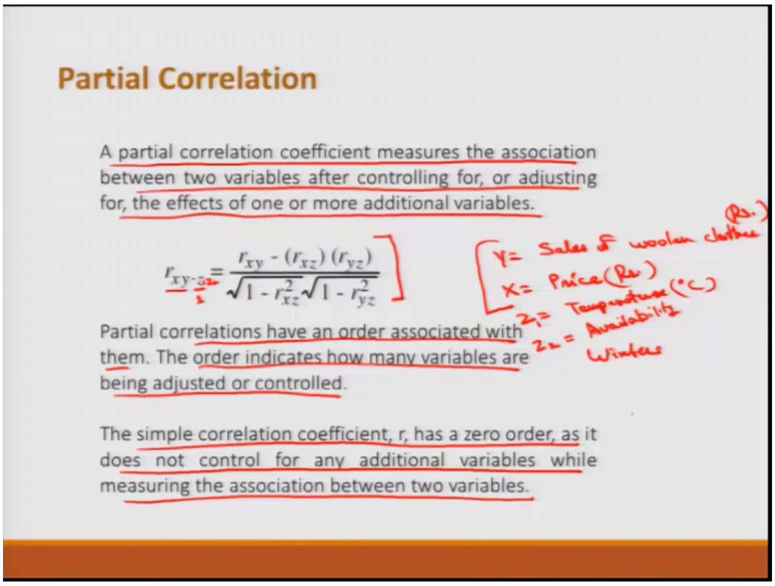
(Refer Slide Time: 15:50)



A Nonlinear Relationship for Which r = 0

So, there is a non-linear relationship as we discussed before here there is a non-linear relationship there is a relationship, but the correlation coefficient is 0. So, this diagram is again explaining that thing.

(Refer Slide Time: 16:05)



Partial Correlation

A partial correlation coefficient measures the association between two variables after controlling for, or adjusting for, the effects of one or more additional variables.

$$r_{xy\cdot z} = \frac{r_{xy} - (r_{xz})(r_{yz})}{\sqrt{1 - r_{xz}^2}\sqrt{1 - r_{yz}^2}}$$

Y= Sales of woolen clothes (Rs.)
X= Price (Rs)
z_1= Temperature (°C)
z_2= Availability
Winters

Partial correlations have an order associated with them. The order indicates how many variables are being adjusted or controlled.

The simple correlation coefficient, r, has a zero order, as it does not control for any additional variables while measuring the association between two variables.

Though next is partial correlation partial a partial correlation coefficient measure the association between two variables after controlling for adjusting for the effects of one or more additional variables, here the partial correlation between x and y while controlling the variable z is given by this relation partial correlations have an order associated with them. So, what is the practical significance here? Let me say I am looking on the sales of garments sales of let me put the sales of woollen garments or woollen clothes. So, sales of the woollen clothes is my Y variable, X variable is the price of the garment the higher would be the price the lower would be the sale this is the kind of known thing ok.

But if I say the temperature or the weather is another variable here. So, when I see the correlation coefficient between or association between the sales of woollen cloths and price; while controlling the variable temperature I would see that the lower temperature that is in winters the sales is high. If I just see the relation between X and Y the price would affect the sales the price would produce the sales.

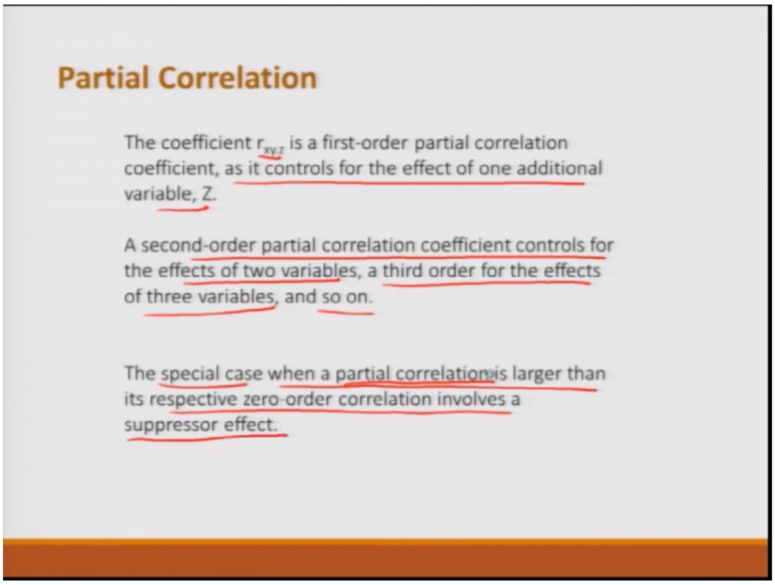
But if temperature another variable is there if it is chilling cold here if the winter season is at peak and the cold is chilling, the sales would be high irrespective of price. So, this is the use of partial correlation; so, here continuous dependent variable would be woollen garment sales measured in rupees continuous independent variable would be the price also measured in rupees. And the single controlled variable that is the single continuous independent variable which we are adjusting for would be the temperature that in measured in degrees rupees this is rupees.

And we may believe that this relationship between woollen garments sales and prices sales go down as prices go up, but it is interesting to note at if this relationship is effected by the temperature that if that is if the relationship is weaker, when we take into account the temperature. Since we suspect customers are more willing to buy woollen garments irrespective of price if it is a cold season.

Partial correlation have an order associated with them the order indicates how many variables are being adjusted or controlled. Simple correlation coefficient r has a 0 order as it does not control for any additional variables while measuring association between two variables. So, this is a one order correlation where controlling one order.

So, let us say if I have also having some other variable Z 2; if this is Z 1 that is availability or market shortage. So, I can put here Z 1 and Z 2 two variables are control that this becomes second order partial correlation.

(Refer Slide Time: 20:05)



Partial Correlation

The coefficient $r_{xy.z}$ is a first-order partial correlation coefficient, as it controls for the effect of one additional variable, Z.

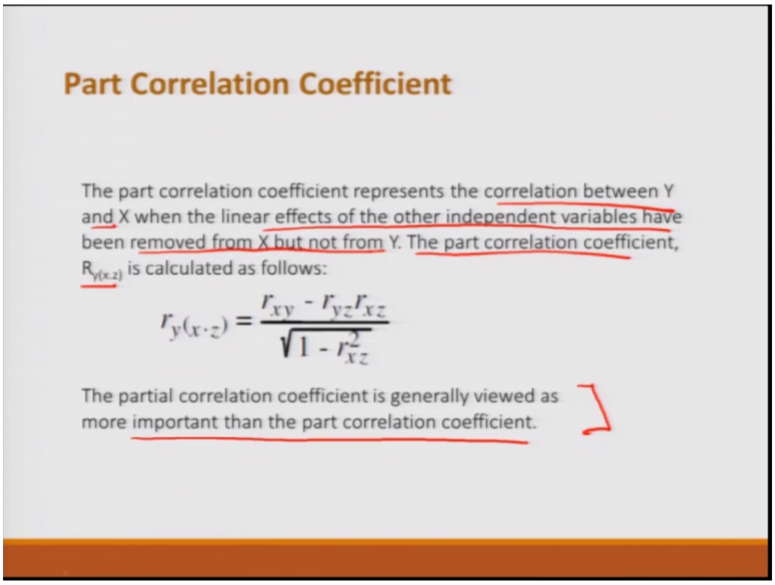
A second-order partial correlation coefficient controls for the effects of two variables, a third order for the effects of three variables, and so on.

The special case when a partial correlation is larger than its respective zero-order correlation involves a suppressor effect.

So, this r xy dot z is a first order partial correlation coefficient, as it controls for the effect of one additional variable. The second order partial correlation coefficient controls for the effect of two variables the third order for the effect of three variables and so, on.

A special case when partial correlation is larger than the respective zero order correlation involves a suppressor effect. Suppressor effect means since we were talking about these two as primary variables, but if Z temperature is effecting the sales so, much that this partial correlation coefficient is higher than this is suppressing my independent variable suppressing my other primary variable price temperature is suppressing the price.

(Refer Slide Time: 21:01)



Part Correlation Coefficient

The part correlation coefficient represents the correlation between Y and X when the linear effects of the other independent variables have been removed from X but not from Y. The part correlation coefficient, $R_{y(x.z)}$ is calculated as follows:

$$r_{y(x\cdot z)} = \frac{r_{xy} - r_{yz}r_{xz}}{\sqrt{1 - r_{xz}^2}}$$

The partial correlation coefficient is generally viewed as more important than the part correlation coefficient.

Next comes in part correlation part correlation coefficient represent the correlation between X and Y; when the linear effects of other independent variables have been removed from X, but not from Y; the part correlation coefficient R y x dot z is calculated as follows. That is the partial correlation coefficient is generally viewed as more important than part correlation coefficient. So, part correlation coefficient is a special case of partial correlation coefficient

So, in partial correlation coefficient if we say whether there were relation between the graduate and undergraduate grades, while we control the IQ. And in part correlation coefficient we can say what is the correlation between the two types of grades controlling the graduate grades for graduate study time. So, this can be one special case; part correlation coefficient can be one special case of partial correlation coefficient. So, with this I would like to stop here.

Thank you.