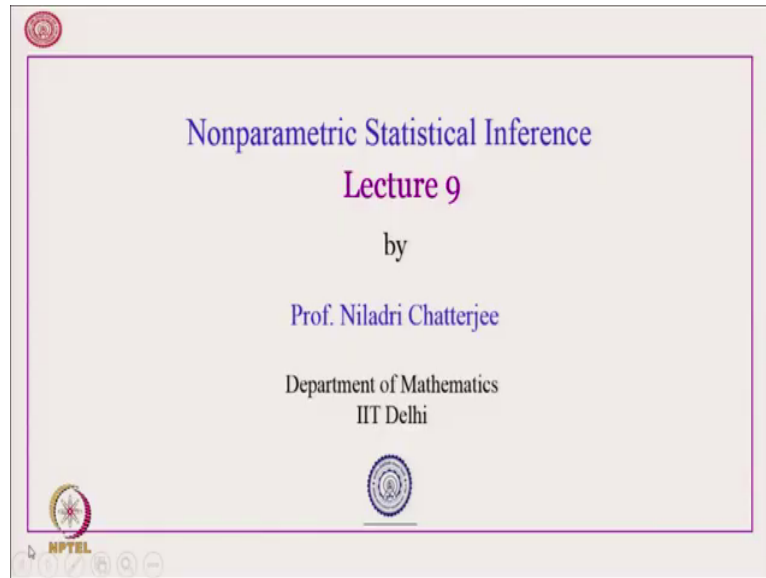


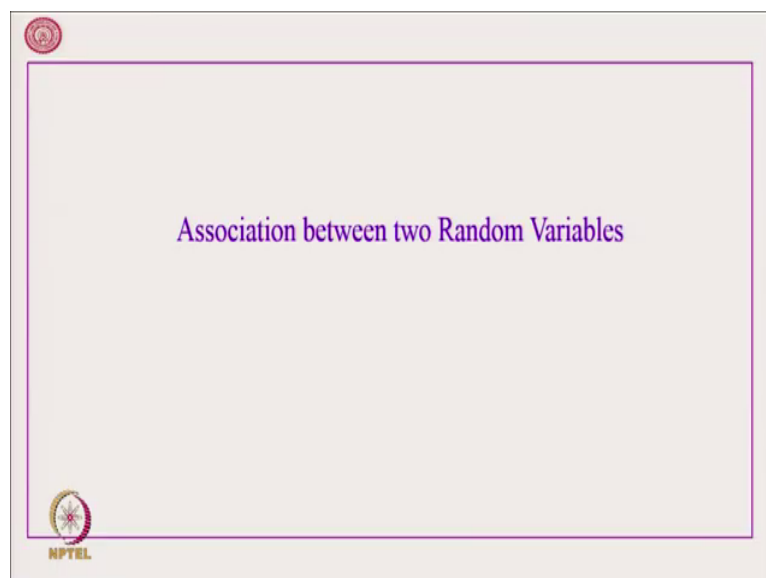
Nonparametric Statistical Inference
Professor Niladri Chatterjee.
Department of Mathematics
Indian Institute of Technology, Delhi
Lecture - 9
Nonparametric Statistical Inference

(Refer Slide Time: 0:18)



Welcome students to the MOOC's series of lectures on nonparametric statistical inference, this is lecture number 9.

(Refer Slide Time: 0:40)



As I said at the end of the last class that in this class we shall study association between two random variables.

(Refer Slide Time: 0:48)

Prologue

In parametric set up we use Covariance as a measure of Association between two variables.

Given X & Y two random variables with a Bivariate probability distribution, their covariance reflects the direction & amount of association or correspondence between the two variables.

Cov(X,Y) can be +ve or -ve and large or small.

Given n pairs of observations $(x_i, y_i), i = 1, 2, \dots, n$

$$Cov(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

However, Cov(X,Y) is NOT unit free.

Handwritten notes:

- If +ve then large X is associated with large Y and small X is associated with small Y.
- If -ve then large X is associated with small Y and small X is associated with large Y.
- If +ve $\Rightarrow$ large values of X are associated with the large values of Y. If -ve $\Rightarrow$ small X associated with small Y.

I hope all of You know how we study the association of two random variables in a parametric setup and the corresponding measure is covariance. So, the situation is like this, given X and Y two random variables, we take bivariate probability distribution, this is very important because if X and Y are independent, then they do not have any covariance, that value is going to be 0. But if there is a joint distribution, which we call bivariate probability distribution, then their covariance reflects the direction and amount of association or correspondence between the two random variables.

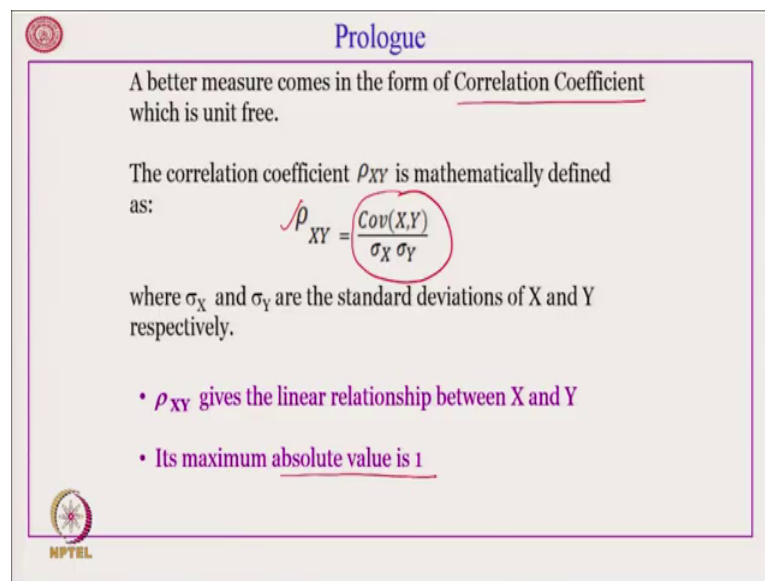
So, we know that covariance of X, Y can be positive or negative, that is direction is plus or minus. If it is plus, implies that large values of X are associated with the large values of Y and similarly, for small values, that is small x is associated with a small y. On the other hand, if it is negative, then we know that if the covariance is negative, then large value of X is associated with small values of Y. And similarly, large values of Y are associated with small values of X and it can be large or small, depending upon how much space is there for X and Y.

And in particular, if $X_1, Y_1, X_2, Y_2, \dots, X_n, Y_n$ is n pairs of observation, the covariance of X, Y is calculated using this following formula. It is

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Therefore, depending upon these magnitudes, the overall value of the covariance will be determined and note that it is not unit free. That means that suppose X and Y are height and weight of boys, then if height is measured in meters and weight is measured in kilograms, then these values will be smaller in comparison with when height is measured in foot and weight is measured in say pounds. And therefore, the magnitude of the covariance may change.

(Refer Slide Time: 4:33)



Prologue

A better measure comes in the form of Correlation Coefficient which is unit free.

The correlation coefficient ρ_{XY} is mathematically defined as:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

where σ_X and σ_Y are the standard deviations of X and Y respectively.

- ρ_{XY} gives the linear relationship between X and Y
- Its maximum absolute value is 1

 NPTEL

Hence, a better measure comes in the form of correlation coefficient, which is unit free and it is typically denoted by the symbol ρ and $\rho_{X,Y}$ is defined as

$$\frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

where σ_X and σ_Y are the standard deviations for X and Y. Note that the $\rho_{X,Y}$ gives a measure of the linear relationship between X and Y and like covariance, it can be plus or minus, but the maximum absolute value is 1.

(Refer Slide Time: 5:19)

Illustration

						Mean	σ^2
X	-2	-1	0	1	2	0 ✓	2 ✓
Y1	-1	0	1	2	3	1 ✓	2 ✓
Y2	4	3	2	1	0	2 ✓	2
Z	5	2	1	2	5	3	11.8 - 9 = 2.8

$n = 5$
(X, Y_1)

$$\rho_{X,Y1} = ((-2+0+0+2+6)/5 - 0*1) / (2^{1/2} * 2^{1/2}) = 1$$

$$\rho_{X,Y2} = ((-8-3+0+1+0)/5 - 0*2) / (2^{1/2} * 2^{1/2}) = -1$$

$$\rho_{X,Z} = ((-10-2+0+2+10)/5 - 0*3) / (2^{1/2} * 2.8^{1/2}) = 0$$

Observe: $Y1 = X + 1$ $Y2 = -X + 2$ $Z = X^2 + 1$

linearly related
if X increases Y decreases

So, let me illustrate. Consider 5 points, that is n is equal to 5 and we have X_1, X_2, X_3, X_4 and X_5 , we are checking its association with three other random variables Y_1, Y_2 , and Z . So, let us consider X and Y_1 that means

You are looking at these pairs of observations $(-2, -1), (-1, 0), (0, 1), (1, 2)$ and $(2, 3)$. Therefore, what is going to be the correlation between X and Y_1 , we look at this table the mean of X is 0 and its variance is 2 the mean of Y_1 is 1 and variance is 2.

Therefore, $\rho_{X,Y1}$ is $(\sum X_i Y_i) / n - \bar{X} * \bar{Y}$, whole divided by $\sigma_X \sigma_Y$. And therefore, it is coming out to be -2 into -1 which is 2, -1 into 0 which is 0, 0 multiplied by 1 which is 0, 1 multiplied by 2 which is 2 and 2 multiplied by 3 which is 6, that divided by 5 minus $\bar{X}$ is equal to 0, $\bar{Y}$ is equal to 1 and their standard deviations are $\sqrt{2}, \sqrt{2}$. Therefore, together we are getting the value is coming out to be 1. Because $2 + 2$ is 4, $4 + 6$ is 10, $10 / 5$ is 2, therefore, $2 / 2 = 1$.

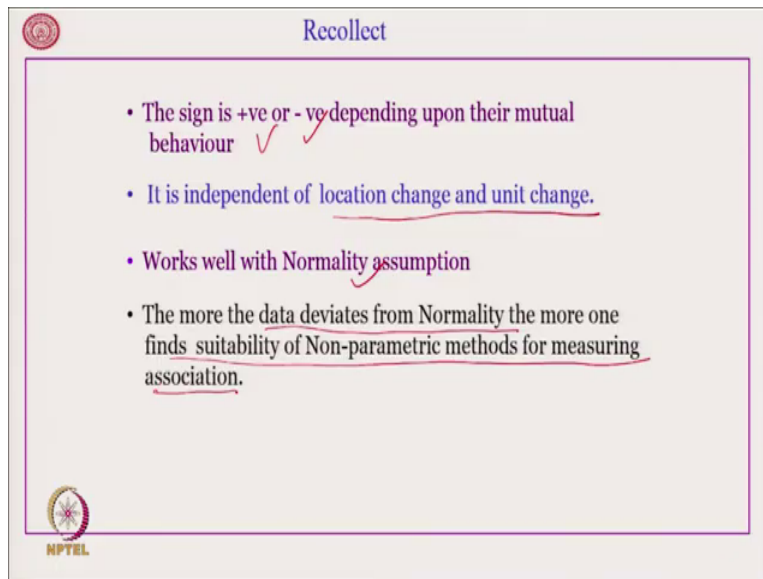
In a similar way, $\rho_{X,Y2}$ coming out to be -1 and $\rho_{X,Z}$ is coming out to be 0. Now, let us look into that data, we can see that they have the following relationship, Y_1 is equal to X plus 1, that means Y and X has a positive relationship as X increases Y_1 increases as X decreases Y_1 decreases and they are perfectly linear. Therefore, the correlation is coming out to be positive and because it is perfectly linear, its value is 1.

On the other hand, if we consider Y_2 , it is nothing but $-X + 2$, therefore for -2 it is coming out to be 4, for -1 it is coming out to be 3 and for 2 it is coming out to be 0. What is there in this data? Firstly, they are linear or linearly related and secondly if X increases then Y decreases.

Because they are perfectly linear, the magnitude of the correlation is 1, but because increasing X means decreasing Y , we get a negative sign there that is it is negatively correlated, that is very clear from this value.

Now, if we consider the correlation of X and Z we can see that their relationship is Z is equal to $X^2 + 1$. Therefore, there is no linearity between them and therefore, we can see that the correlation between them is 0. So, this is some intuition that we need to know when we talk about the association of 2 random variables.

(Refer Slide Time: 9:52)



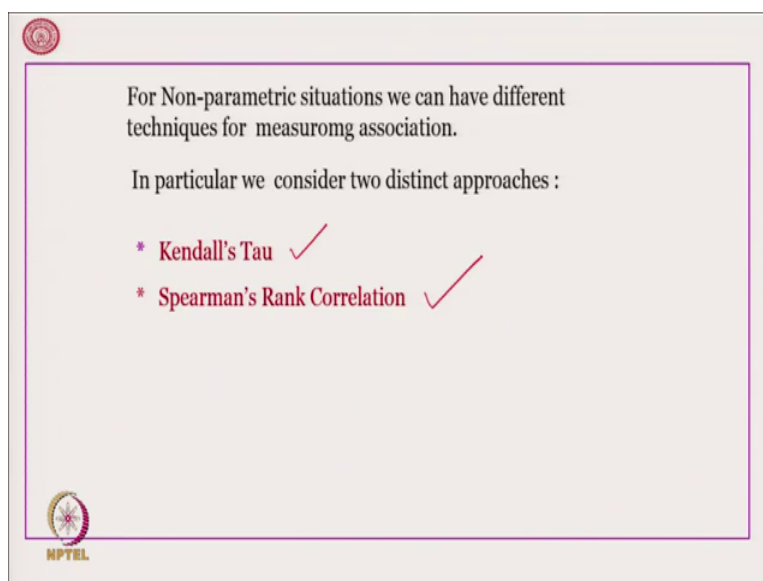
Recollect

- The sign is +ve or -ve depending upon their mutual behaviour ✓ ✓
- It is independent of location change and unit change.
- Works well with Normality assumption
- The more the data deviates from Normality the more one finds suitability of Non-parametric methods for measuring association.

NPTEL

Therefore, let us recollect that the sign is positive or negative depending upon their mutual behavior, but it is independent of location change and unit change, it works well with normality assumption. As we said before, that the parametric statistical inference is based on a normality assumption. Therefore, the correlation coefficient that we discussed works well under normality assumption. However, the more the data deviates from normality, the more one finds the suitability of nonparametric methods for measuring Association. That is, if the data is non normal, then rather we should look at non parametric way of measuring their associations.

(Refer Slide Time: 10:48)



For Non-parametric situations we can have different techniques for measuring association.

In particular we consider two distinct approaches :

- * Kendall's Tau ✓
- * Spearman's Rank Correlation ✓

NPTEL

So, in this class, we should study two important non parametric tests for association, one is called Kendall's Tau and Spearman's Rank correlation.

(Refer Slide Time: 11:05)


Recollect

- The sign is +ve or - ve depending upon their mutual behaviour ✓✓
- It is independent of location change and unit change.
- Works well with Normality assumption
- The more the data deviates from Normality the more one finds suitability of Non-parametric methods for measuring association.


NPTEL

So, let us first look into Kendall's Tau, often it is written like this from the Greek alphabet τ .

(Refer Slide Time: 11:10)



Measures the association between X and Y from a bi-variate distribution based on n observations

$$\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$$

They can be Numeral, Ordinal even Rank data etc.

Example:

(Numeral, Numeral):	$\{(50, 57), (80.5, 85), (67.5, 54.5), (76, 75)\}$
(Numeral, Symbolic):	$\{(50, C), (80.5, A), (67.5, D), (76, B)\}$
(Rank, Rank) :	$\{(4, 3), (1, 1), (3, 4), (2, 2)\}$
(Rank, Grade) :	$\{(4, C), (1, A), (3, D), (2, B)\}$


NPTEL

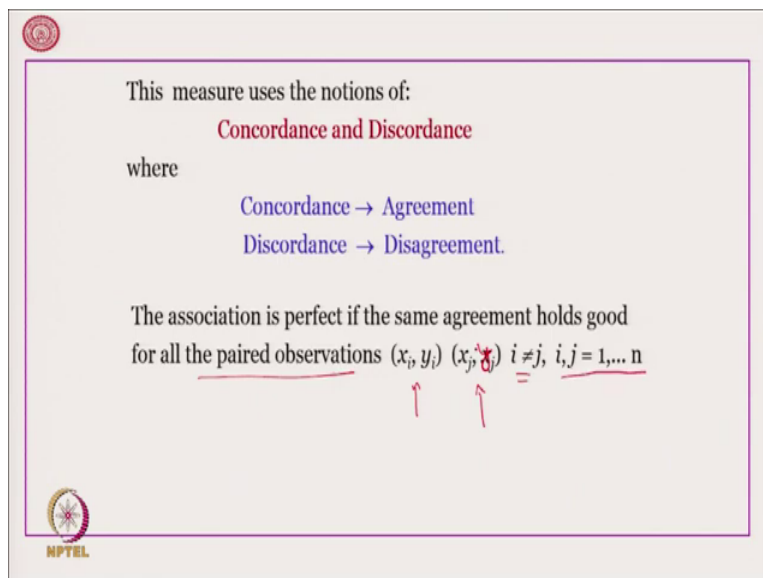
So, what it is? It measures the association between X and Y from a bivariate distribution based on n observations. So, let X_1, Y_1, X_2, Y_2 , up to X_n, Y_n be the n pairs of observation taken from the bivariate distribution. They can be numeral, they can be ordinal or they can be even rank data.

Say for example, we are looking at marks of 4 students in 2 subjects. So, one way of looking at them is both of them are numerals. So, first student got 50 in one subject and 57 in the other subject, second subject got 80.5 and 85. Similarly, 67.5 and 54.5 for the third student and these 2 are for the fourth student.

Again, one of them can be numeral, but the other can be symbolic. Many institutions they give grades instead of marks and suppose the grades obtained by that four students in the second subject are C, A, D and B respectively. One can even represent that data using ranks, that means with respect to the first subject, the ranks of the people for students are fourth, first, third, and second, as we can see that this is the highest one, this is the second highest one, this is the third highest and this is the fourth highest.

And similarly, the ranks for the second subject they have given below. Or the data can be something where one is rank, other one is symbolic. In short, we want to say that the data can be in many different forms, they need not be numeric. And therefore, in such cases, the correlation coefficient that we discussed cannot be computed, because they need arithmetic manipulation of data with the help of the formula that we have given earlier.

(Refer Slide Time: 13:35)



This measure uses the notions of:

Concordance and Discordance

where

Concordance → Agreement
Discordance → Disagreement.

The association is perfect if the same agreement holds good for all the paired observations (x_i, y_i) and (x_j, y_j) $i \neq j, i, j = 1, \dots, n$

The slide includes a logo in the top left corner and the NPTEL logo in the bottom left corner. Red arrows point to the (x_i, y_i) and (x_j, y_j) terms in the text below.

Then the question comes how to measure their association. In Kendall's Tau, it uses 2 notions, one is called concordance or that is called discordance, where concordance means agreement and discordance means disagreement. So, what does it mean? It means that the association is perfect for the i th and j th data point namely (x_i, y_i) and (x_j, y_j) for all paired observations when $i \neq j$ and i, j is equal to from 1 to n .

(Refer Slide Time: 14:25)



Definition:
A relation is said to be in perfect concordance if for any two Pairs (X_i, Y_i) and (X_j, Y_j)

- * whenever $X_i < X_j$ we also have $Y_i < Y_j$ or
- * whenever $X_i > X_j$ we also have $Y_i > Y_j$

Or equivalently:
$$(X_i - X_j) * (Y_i - Y_j) > 0$$

Example: Consider the set of 5 observations (X_i, Y_i) : $i = 1, \dots, 5$

$(-10) * (3-12) > 0$

$\{(1, 3), (10, 12), (8, 7), (7, 5), (14, 18)\}$

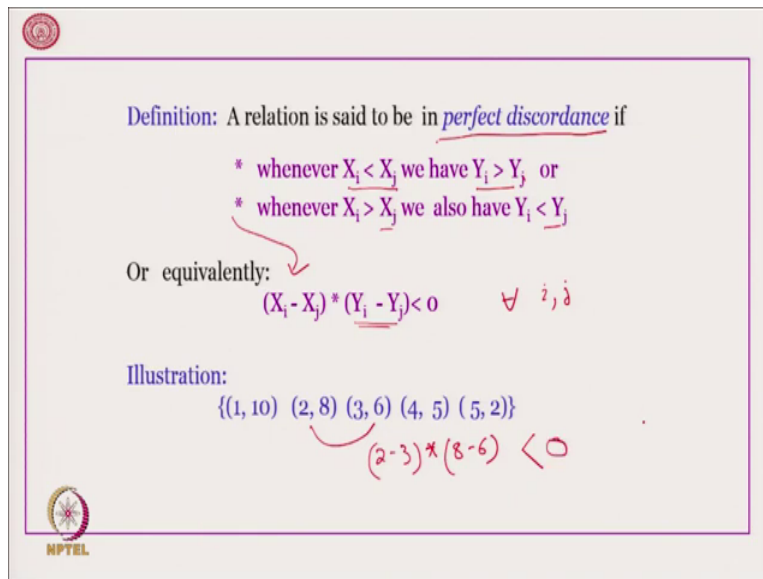
Shows Perfect concordance.



So, by definition, a relation is said to be in perfect concordance if for any two pairs X_i, Y_i and X_j, Y_j , whenever, $X_i < X_j$, we also have $Y_i < Y_j$ or whenever, $X_i > X_j$, we also have $Y_i > Y_j$. Together we can write that $X_i - X_j$ multiplied by $Y_i - Y_j > 0$. For example, consider the set of 5 observations, X_i, Y_i , i is equal to 1 to 5, (1, 3), (10, 12), (8, 7), (7, 5), (14, 18). Then we can see that it shows perfect concordance.

For example, if we consider these 2 pairs, here the first observation the value is increased and the second observation the value is also increased. Therefore, we are looking at 1 minus 10 multiplied by 3 minus 12. And we can see that this is going to be greater than 0. Similarly, we can check with respect to all the possible pairs, that is 5C_2 many pairs and we can see that for all of them, this value is coming out to be greater than 0 and therefore, this is a perfect concordance.

(Refer Slide Time: 16:11)



Definition: A relation is said to be in perfect discordance if

- * whenever $X_i < X_j$ we have $Y_i > Y_j$ or
- * whenever $X_i > X_j$ we also have $Y_i < Y_j$

Or equivalently:

$$(X_i - X_j) * (Y_i - Y_j) < 0 \quad \forall i, j$$

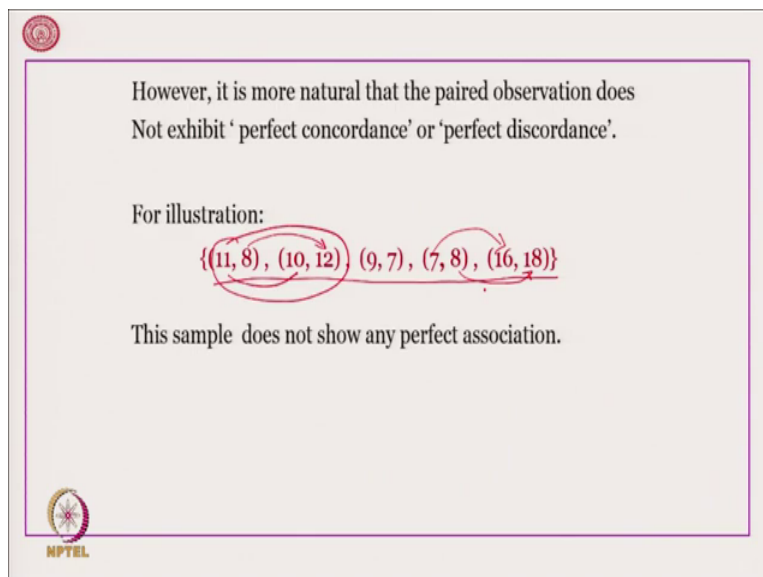
Illustration:

$\{(1, 10) (2, 8) (3, 6) (4, 5) (5, 2)\}$

$(2-3) * (8-6) < 0$

At the same time, a relation is said to be perfect discordance if whenever $X_i < X_j$, we have $Y_i > Y_j$, that is, if the value of X is increased, the value of Y is decreased or if value of X is decreased, then value of Y is increased. In other words, these two together we can write as $X_i - X_j$ multiplied by $Y_i - Y_j$ has to be less than 0 for all i, j. For example, let us choose arbitrarily these two pairs, therefore, we get 2 - 3 multiplied by 8 - 6 and we know that this is coming out to be less than 0. The same one can compute with respect to all the 5C_2 many pairs.

(Refer Slide Time: 17:20)



However, it is more natural that the paired observation does Not exhibit 'perfect concordance' or 'perfect discordance'.

For illustration:

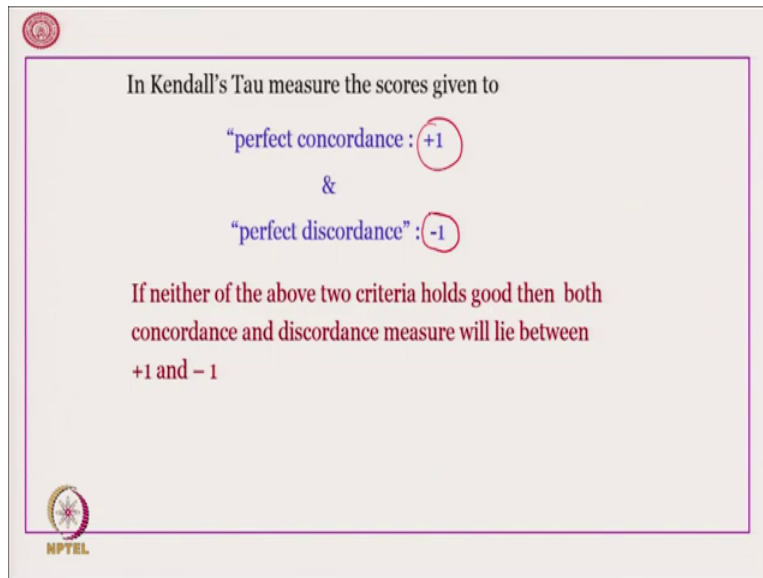
$\{(11, 8), (10, 12), (9, 7), (7, 8), (16, 18)\}$

This sample does not show any perfect association.

However, it is more natural that the paired observation does not exhibit perfect concordance or perfect discordance with respect to all the pairs. Again, for illustration consider these 5

tuples. If we consider these two, (11, 8) and (10, 12), we can see that value of X is decreased, but value of Y is increased therefore they are in discordance. However, if we consider these two pairs, the value of X is increased and value of Y is also increased. Therefore, there is a concordance between these two pairs. Therefore, this sample does not show any perfect association. And that is very common, that is what in general we would expect.

(Refer Slide Time: 18:30)



In Kendall's Tau measure the scores given to

“perfect concordance” : +1

&

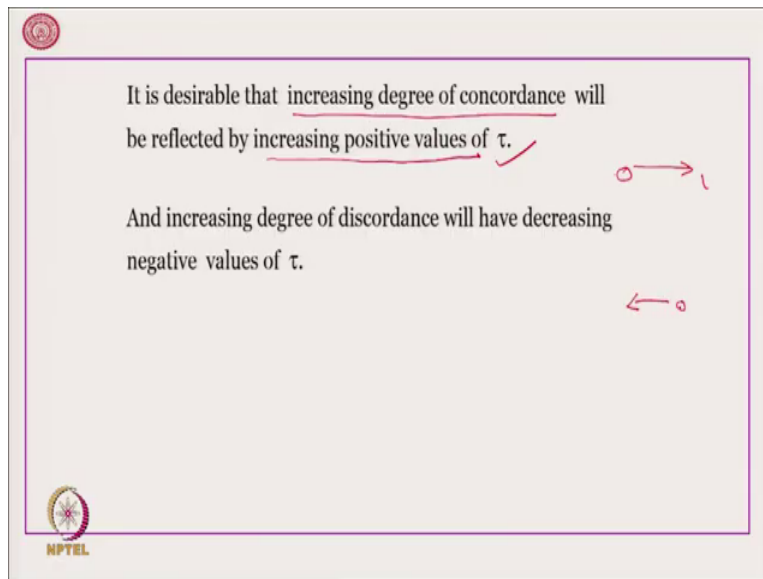
“perfect discordance” : -1

If neither of the above two criteria holds good then both concordance and discordance measure will lie between +1 and -1

NPTEL

And therefore, we need to give a measure which resembles a degree of association between them. So, in Kendall's Tau, the scores are given as follows. If it is perfectly concordant pair or they are in perfect concordance, then the value given will be plus 1, if perfect discordance then value will be given minus 1, if neither of the above two criteria holds good, then both concordance and discordant measure will lie between plus 1 and minus 1. Thus, we can see a similarity with respect to correlation coefficient.

(Refer Slide Time: 19:13)



It is desirable that increasing degree of concordance will be reflected by increasing positive values of τ . ✓

And increasing degree of discordance will have decreasing negative values of τ .

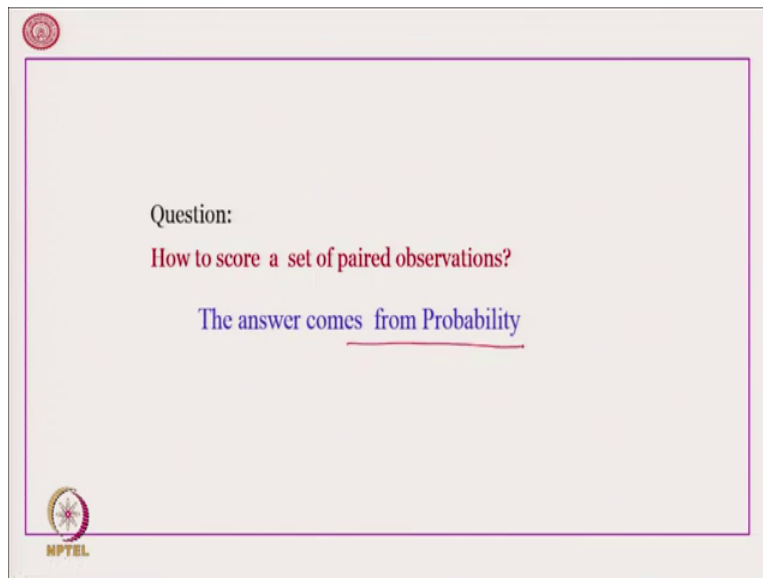
0 → 1

← 0

NPTEL

It is desirable that increasing degree of concordance will be reflected by increasing positive values of τ . That means, the more is the degree of concordance then the value of τ will go from 0 to 1. And similarly, increasing degree of discordance will be shown by the negative value as it is going from 0 to -1.

(Refer Slide Time: 19:44)



Question:

How to score a set of paired observations?

The answer comes from Probability

NPTEL

But the main question is still there. How to score a set of paired observations? We understood what is going to be for perfect concordance or perfect discordance, but how to score in between? this answer comes from probability.

(Refer Slide Time: 20:04)

Let P_c and P_d denote the probabilities of perfect concordance and discordance, respectively, for any two randomly chosen pairs i & j .

This is to be calculated by taking into consideration all the $\binom{n}{2}$ many pairs of observations (X_i, Y_i) and (X_j, Y_j) , where $i \neq j$, and $i, j \in \{1, 2, \dots, n\}$

$\binom{n}{2}$



So, let P_c and P_d denote the probabilities of perfect concordance and perfect discordance respectively, for any two randomly chosen pairs i, j . Now, this is to be calculated by taking into consideration all the ${}^n C_2$ many pairs (X_i, Y_i) and (X_j, Y_j) , where $i \neq j$ and they are running from 1 to n . Or in other words, what we are saying we will consider all the ${}^n C_2$ many pairs and we shall try to see how many of them are concordant and how many of them are discordant. So, from there, we shall try to get a measure of the probability of concordance and discordance.

(Refer Slide Time: 21:05)

Definition: Given $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ ✓

$\tau = \frac{\text{Pr (Perfect Concordance)} - \text{Pr(Perfect Discordance)}}{1}$

$(\tau = P_c - P_d)$

for any two arbitrarily chosen pairs.

Hence

$\tau = \text{Pr}((X_i - X_j) * (Y_i - Y_j) > 0) - \text{Pr}((X_i - X_j) * (Y_i - Y_j) < 0)$

for an arbitrarily selected pair (X_i, Y_i) and (X_j, Y_j)

Assumption:
the marginal distributions of X and Y are continuous



So, given X_1, Y_1, X_2, Y_2 and X_n, Y_n that tau value or Kendall's τ is probability of perfect concordance minus probability of perfect discordance. Or in short, we write tau is equal to P_c

minus P_d where P_c is the probability of perfect concordance and P_d is the probability of perfect discordance. Hence, Tau is probability $X_i - X_j$ multiplied by $Y_i - Y_j$ is greater than 0 minus probability $X_i - X_j$ multiplied by $Y_i - Y_j < 0$.

As we have already discussed that if these two pairs are discordant, then this product is going to be less than 0. And if they are concordant, then this product is going to be greater than 0. And this we are telling for an arbitrarily selected pair X_i, Y_i and X_j, Y_j . However, Kendall's Tau has a basic assumption that we have seen with respect to most of our nonparametric tests that the marginal distributions of X and Y are going to be continuous.

(Refer Slide Time: 22:29)

$$\begin{aligned}
 &\text{Now } P_c \\
 &= \Pr((X_i - X_j) * (Y_i - Y_j) > 0) \\
 &= \Pr([(X_i > X_j) \cap (Y_i > Y_j)] \cup [(X_i < X_j) \cap (Y_i < Y_j)]) \\
 &= \Pr((X_i > X_j) \cap (Y_i > Y_j)) + \Pr((X_i < X_j) \cap (Y_i < Y_j)) \\
 &= \Pr(X_i > X_j) - \Pr((X_i > X_j) \cap (Y_i < Y_j)) + \Pr(X_i < X_j) - \Pr((X_i < X_j) \cap (Y_i > Y_j)) \\
 &= \Pr(X_i > X_j) + \Pr(X_i < X_j) - [\Pr((X_i > X_j) \cap (Y_i < Y_j)) + \Pr((X_i < X_j) \cap (Y_i > Y_j))] \\
 &= 1 - P_d
 \end{aligned}$$

Now, what is P_c , it is the probability of perfect concordance. Therefore, it is the probability that $X_i - X_j$ multiplied by $Y_i - Y_j > 0$. That means, it is the probability that $X_i > X_j$ and $Y_i > Y_j$ union $X_i < X_j$ and $Y_i < Y_j$ that means, both of them will give positive or both of them will give us negative values, so that the product will remain positive.

Since these two are disjoint events, we can write as probability $X_i > X_j$ and Y_i greater than Y_j plus probability $X_i < X_j$ and Y_i less than Y_j . Now, this event X_i greater than X_j and $Y_i > Y_j$ can be written as probability $X_i > X_j$ minus probability $X_i > X_j$ and $Y_i < Y_j$. This is very straightforward, and we can see it very easily from the Venn diagram. So, I am not going to explain that, you convince yourself that probability of this event can be written as the difference of these two probabilities.

And in a very similar way, probability $X_i < X_j$ and $Y_i < Y_j$, this can be written as probability $X_i < X_j$ minus probability $X_i < X_j$ and $Y_i > Y_j$, here it was less than now, it becomes greater

than. Therefore, this whole thing can be written as probability $X_i > X_j$ plus probability that $X_i < X_j$ minus probability $X_i > X_j$ and $Y_i < Y_j$ plus because minus we have taken common probability X_i , X_j and $Y_i > Y_j$.

Now, this part is equal to 1, because under continuity, either $X_i > X_j$ or $X_i < X_j$, therefore, their sum is going to be 1. And this is nothing but the probability of discordance between X_i , Y_i and X_j , Y_j . Therefore, we can see that probability of concordance is equal to 1 minus probability of discordance.

(Refer Slide Time: 25:25)



$$\begin{aligned}
 &\text{Now } P_c \\
 &= \Pr((X_i - X_j) * (Y_i - Y_j) > 0) \\
 &= \Pr(((X_i > X_j) \cap (Y_i > Y_j)) \cup ((X_i < X_j) \cap (Y_i < Y_j))) \\
 &= \Pr((X_i > X_j) \cap (Y_i > Y_j)) + \Pr((X_i < X_j) \cap (Y_i < Y_j)) \\
 &= \Pr(X_i > X_j) - \Pr((X_i > X_j) \cap (Y_i < Y_j)) + \\
 &\quad \Pr(X_i < X_j) - \Pr((X_i < X_j) \cap (Y_i > Y_j)) \\
 &= \Pr(X_i > X_j) + \Pr(X_i < X_j) - [\Pr((X_i > X_j) \cap (Y_i < Y_j)) + \\
 &\quad \Pr((X_i < X_j) \cap (Y_i > Y_j))] \\
 &= 1 - P_d
 \end{aligned}$$

Therefore $\tau = P_c - P_d = 1 - 2P_d = 2P_c - 1$



Therefore, that tau or Kendall's τ can be written as P_c minus P_d , which we have already defined can also be written as $1 - 2 * P_d$ or $2 * P_c - 1$. So, these are different formulae for computing Kendall's τ .

(Refer Slide Time: 25:51)

Properties of τ

1) Value of τ is 0 if X and Y are independent.

Suppose X and Y are continuous and independent

Therefore, given i and j $P(X_i > X_j) = P(X_i < X_j)$.
 $P(Y_i > Y_j) = P(Y_i < Y_j)$.

Therefore:

$$P_c = P(X_i > X_j) * P(Y_i > Y_j) + P(X_i < X_j) * P(Y_i < Y_j)$$
$$= P(X_i > X_j) * P(Y_i < Y_j) + P(X_i < X_j) * P(Y_i > Y_j) = P_d$$

Therefore,

$$\tau = P_c - P_d = 0$$

So, let us first examine a few properties of tau. The value of tau is 0 if X and Y are independent. Because X and Y are continuous and independent, therefore, given i and j probability $X_i > X_j$ same as $X_i < X_j$ and similarly, probability $Y_i > Y_j$ is same as probability $Y_i < Y_j$. Therefore, probability of concordance is equal to probability X_i greater than 0 multiplied by probability $Y_i > Y_j$ plus probability $X_i < X_j$ multiplied by probability $Y_i < Y_j$.

Now, because of this equality, we can write this as probability $Y_i < Y_j$ and similarly, this as probability $Y_i > Y_j$. Therefore, we find that P_c is same as probability $X_i > X_j$ multiplied by $Y_i < Y_j$ plus probability $X_i < X_j$ multiplied by probability $Y_i > Y_j$. And we know that this entire thing gives us the probability of discordance. Therefore, if X and Y are independent, probability of concordance is same as probability of discordance. Therefore, the value of tau is equal to P_c minus P_d is equal to 0. Therefore, if X and Y are independent, the value of tau is 0.

(Refer Slide Time: 27:41)

Properties of τ

2) Tau (τ) is symmetric with respect to X & Y
i.e. $\tau(X, Y) = \tau(Y, X)$ — Obvious

3) $\tau(X, Y) = \tau(-X, -Y) = \tau(-Y, -X)$ $x_i > x_j \rightarrow -x_i < -x_j$
 $y_i > y_j \rightarrow -y_i < -y_j$

4) $\tau(X, Y) = -\tau(X, -Y) = -\tau(Y, -X)$ $(x_i - x_j)(y_i - y_j) = +ve$
 $(x_i - x_j)(-y_i + y_j) = -ve$
 $(-x_i + x_j)(-y_i + y_j) = +ve$

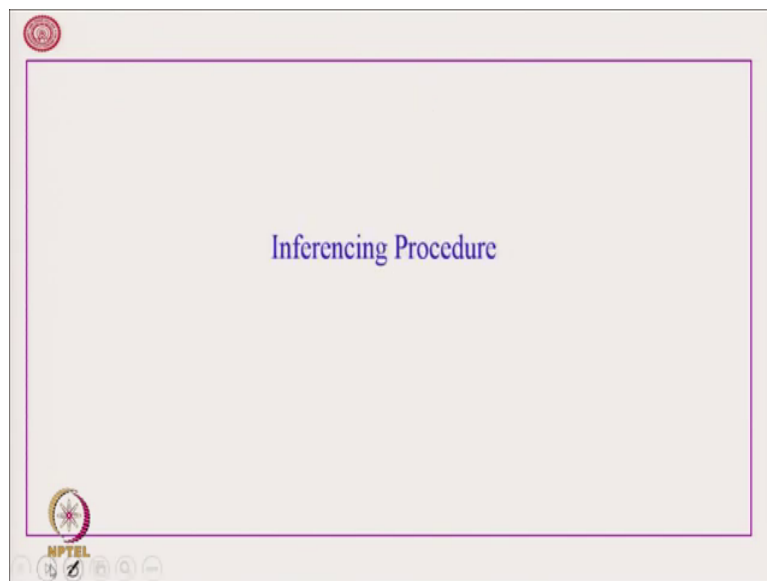
The measure will be invariant under all transformations of X and Y for which order of magnitude is preserved. $\Rightarrow (-x_i + x_j)(-y_i + y_j) = +ve$

Let us examine some other properties of tau. Tau is symmetric with respect to X and Y. That is $\tau(X, Y) = \tau(Y, X)$, which is pretty obvious, I am not going to prove anything here. You can easily convince yourself $\tau(X, Y) = \tau(-X, -Y)$, this is again very obvious, because we are considering the negative values of X and negative values of Y.

Therefore, the relative values if X_i is greater than X_j will now change to $-X_i$ is less than $-X_j$ and similarly, Y_i greater than Y_j implies $-Y_i$ is less than $-Y_j$ and therefore, when we are taking the product whenever this was positive then $-X_i$ plus X_j multiplied by $-Y_i$ plus Y_j will also be positive. Similarly, with respect to negative, therefore, $\tau(X, Y) = \tau(-X, -Y)$, which is again by this observation is $\tau(-Y, -X)$.

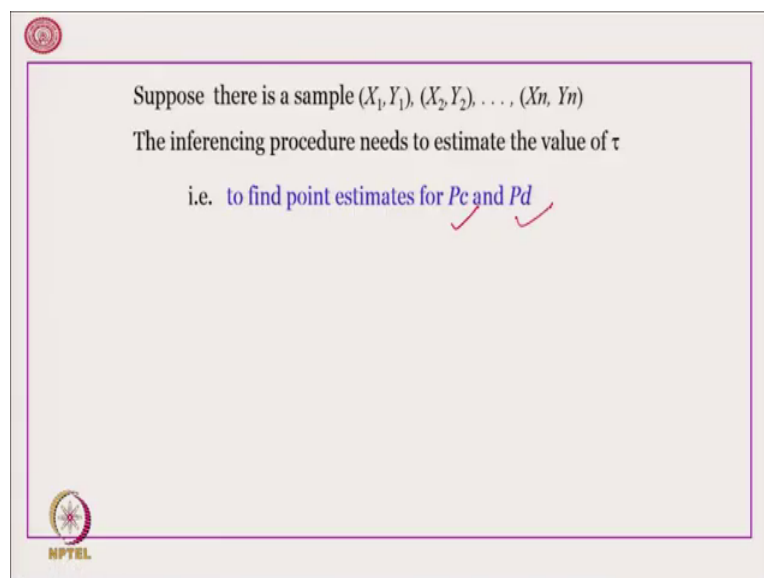
In a similar way, $\tau(X, Y)$ is $-\tau(X, -Y)$ or it is equal to $-\tau(Y, -X)$. Or in other words, the measure will be invariant and that all transformations of X and Y for which order of magnitude is preserved. Because, we are not looking at the exact value, we are looking at whether it is positive or negative. Therefore, if the relative orders are preserved, then there is no change in the value of the Kendall's Tau.

(Refer Slide Time: 30:17)



Inferencing procedure.

(Refer Slide Time: 30:19)



Suppose, there is a sample $(X_1, Y_1), (X_2, Y_2), (X_n, Y_n)$ the inferencing procedure needs to estimate the value of τ . That is to find point estimates for P_c and P_d , that is probability of concordance and probability of discordance.

(Refer Slide Time: 30:40)

Procedure

For each pair i, j where $i \neq j$, and $i, j \in \{1, 2, \dots, n\}$ define

$$A_{ij} = \text{sgn}(X_i - X_j) * \text{sgn}(Y_i - Y_j),$$

where sgn is the sign function

Then A_{ij} can be defined as follows:

$$A_{ij} = \begin{cases} 1 & \text{if the pair is Concordant} \\ -1 & \text{if the pair is Discordant} \\ 0 & \text{Otherwise} \end{cases}$$

So, for each pair i, j , when $i \neq j$, define A_{ij} is equal to sgn of $X_i - X_j$ multiplied by sgn of $Y_i - Y_j$ for sgn is the sign function that means, it is +1 or -1 depending upon whether it is positive or negative. Then A_{ij} can be defined as follows, it is +1 if the pair is concordant, it is -1 if the pair is discordant, and it is 0 otherwise. That means, when a pair is neither concordant nor discordant and since it is a pair, it means that there will be some equality.

Although under continuity assumption, equalities should not happen, but in reality when we collect data, discrete data, there may be equalities, therefore, we need to exclude those data points from further calculation.

(Refer Slide Time: 31:49)

The marginal probability distribution of A_{ij} 's is

$$f_{i,j}(a_{i,j}) = \begin{cases} P_c & \text{if } a_{i,j} = 1 \\ P_d & \text{if } a_{i,j} = -1 \\ 1 - P_c - P_d & \text{if } a_{i,j} = 0 \end{cases}$$

Therefore,

$$E(A_{ij}) = 1 \cdot P_c - 1 \cdot P_d = \tau$$

Therefore, the marginal probability distribution of A_{ij} is this, it is P_c , if A_{ij} is equal to 1. That means, probability that A_{ij} will take value 1 that probability is probability of concordance that is P_c , A_{ij} will take a value minus 1 that probability is P_d , and if A_{ij} is equal to 0, that value is $1 - (P_c - P_d)$. Therefore, expected value of A_{ij} is equal to 1 times P_c minus 1 times P_d plus 0 times this quantity which we can ignore or in other words, what we have is 1 times P_c minus 1 times P_d , which is equal to the value of τ . Therefore, this A_{ij} helps us in estimating the value of τ .

(Refer Slide Time: 32:43)



$$\sum_{1 \leq i < j \leq n} \sum A_{ij}$$

= number of concordant pairs - number of discordant pairs.

In case of no ties in X and Y values,
a pair is either concordant or discordant ✓ ✓

Therefore

$$\frac{\text{number of discordant pairs}}{= \binom{n}{2} - \text{number of concordant pairs}}$$


Now, if we take summation over A_{ij} the number of concordant pairs minus number of discordant pairs that is what we are getting. If there is no ties, each pair will be either concordant or discordant as I have mentioned. Therefore, the number of discordant pairs is going to be nC_2 minus the number of concordant pairs.

(Refer Slide Time: 33:10)

Hence an unbiased estimator for τ is provided by

$$T = \sum_{1 \leq i < j \leq n} \sum_{\binom{n}{2}} \frac{A_{ij}}{n(n-1)} = 2 \sum_{1 \leq i < j \leq n} \frac{A_{ij}}{n(n-1)}$$

Thus T gives us the measure of association.

Illustration:

Consider $\{(1, 15), (6, 12), (4, 10), (8, 8), (6, 9)\}$

Hence, an unbiased estimator for tau is provided by sigma over i and j, i, j going from 1 to n and i less than j, A_{ij} divided by nC_2 , because there are nC_2 many possible pairs, which is coming out to be

$$2 \sum_{1 \leq i < j \leq n} \frac{A_{ij}}{n(n-1)}$$

So, this T value we compute from the sample and it gives a measure of association between X and Y. For illustration, consider these pairs (1,15), (6, 12), (4, 10), (8, 8) and (6, 9).

(Refer Slide Time: 33:57)

Illustration

There are 10 pairs. ✓
For each of them we calculate A_{ij}:
 $\{(1, 15), (6, 12), (4, 10), (8, 8), (6, 9)\}$

i	j	A _{ij}
1	2 ✓	-1 ✓
1	3 ✓	-1
1	4 ✓	-1
1	5 ✓	-1
2	3 ✓	+1 ✓
2	4 ✓	-1
2	5 ✓	0
3	4 ✓	-1
3	5 ✓	-1
4	5 ✓	-1

Since, there are 5 observations, therefore, we have 5C_2 that means 10 pairs. So, let us consider these 10 pair (1, 2), (1, 3), (1, 4), (1, 5), (2, 3), (2, 4), (2, 5), (3, 4), (3, 5) and (4, 5). Now, we have to check how many of them are concordant, how many of them are discordant, so, we go pairwise. So, 1 and 5 as we can see that has 1, X values are 1 and 6 and 15 and 12 therefore, this is a discordant pair, therefore, the value of A_{ij} is equal to minus 1. Similarly, You can calculate for all others, let us just check the pair 2 and 3, it is 6, 4 and 12, 10.

That means, as I go from second sample to the third sample, value of X is decreasing and value of Y is decreasing therefore, that is a concordant pair. And therefore, as we see, we have got a plus 1 there. What is with respect to 2 and 5? Since the X values are 6, for both of them, therefore, this pair is neither concordant nor discordant, therefore, we get the value 0. So similarly you can calculate for all other pairs, I am not going to do that.

(Refer Slide Time: 35:32)

Illustration

There are 10 pairs.

For each of them we calculate A_{ij} :

$\{(1, 15), (6, 12), (4, 10), (8, 8), (6, 9)\}$

$P_c = 1/9$ ✓

$P_d = 8/9$ ✓

Therefore computed value of τ
 $= -7/9$ (T)

i	j	A_{ij}
1	2	-1
1	3	-1
1	4	-1
1	5	-1
2	3	+1
2	4	-1
2	5	0
3	4	-1
3	5	-1
4	5	-1

But now, what is see that we have to discard this pair, because it is neither concordant nor discordant and then we find that there is only 1 pair, which are concordant and remaining 8 pairs, they are all discordant. Therefore, probability of concordance is equal to $1/9$, probability of discordance is equal to $8/9$. Therefore, the computed value of tau from the sample is $-7/9$. So, that is the estimate of the Kendall's Tau for this sample.

(Refer Slide Time: 36:14)

Illustration 1

Consider
 $\{(5, 15), (8, 14), (6, 12), (7, 18)\}$

Therefore we have the following:

i	j	A_{ij}
1	2	-1
1	3	-1
1	4	+1
2	3	+1
2	4	-1
3	4	+1

Now Consider them with
A different order:
 $\{(8, 14), (7, 18), (6, 12), (5, 15)\}$

Hence we have:

i	j	A_{ij}
1	2	-1
1	3	+1
1	4	-1
2	3	+1
2	4	+1
3	4	-1

Thus Distribution of A_{ij} does not change with permutation

Now, let us see a few things considered this pair (5, 15), (8, 14), (6, 12) and (7, 18). Therefore, we have the following pairs 4C_2 which is 6 and in fact, we can see that (1, 2), (1, 3), (1, 4), (2, 3), (2, 4) and (3, 4) and these are the values of A_{ij} s. Now, consider them with a different order, the same pair of observations, but we have permuted them differently, that means, the sample might have been taken in a different order. But what we see that the number of positive and number of negative still remains same. Or in other words, we can say that, the distribution of A_{ij} does not change with permutation. That is an important property of Kendall's Tau, that if You shuffle the data, that does not alter the value of the measure of association that is tau.

(Refer Slide Time: 37:22)

Illustration 2

Consider
 $\{(5, 15), (8, 14), (6, 12), (7, 18)\}$

Therefore we have the following:

i	j	A_{ij}
1	2	-1
1	3	-1
1	4	+1
2	3	+1
2	4	-1
3	4	+1

Now Consider X values replaced
with their ranks

Hence we have:
 $\{(1, 15), (4, 14), (2, 12), (3, 18)\}$

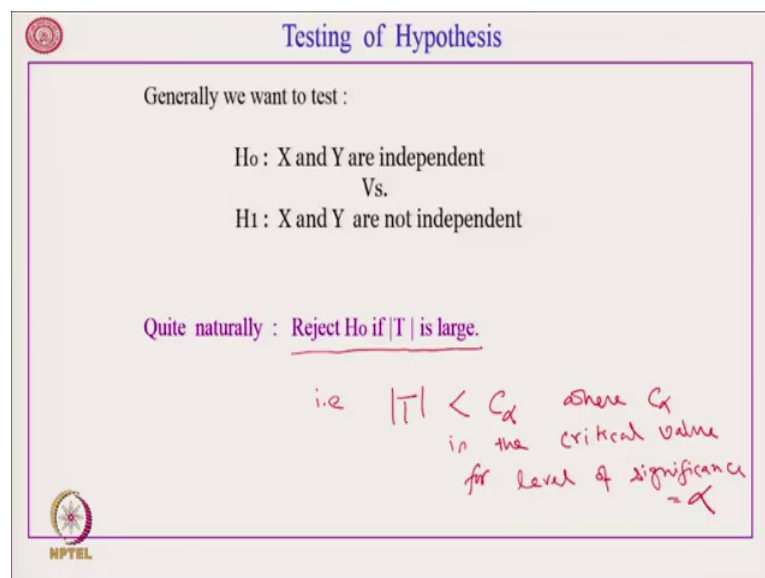
i	j	A_{ij}
1	2	-1
1	3	-1
1	4	+1
2	3	+1
2	4	-1
3	4	-1

Distribution of A_{ij} does not change with Values replaced by Ranks

Let us consider another example. Again, we have taken this 4 pair of observation, we have computed this table of A_{ij} . Now, consider X values are replaced with their ranks. That means, if we look at these the values are 5, 8, 6 and 7. So, suppose we rank them as 1, 2, 3 and 4 then now my data is transformed into this. Therefore, again if we calculate A_{ij} , we find that there is no change in the computation of A_{ij} between these 2 set of data. So, what does it tell us?

It tells us that the distribution of A_{ij} does not change with the values replaced by ranks. What is the advantage? The advantage is that instead of numeric value, if we add A , B , C , etc and they can be ordered perfectly, then there is no change in the value of the measure of association namely Kendall's Tau.

(Refer Slide Time: 38:42)



Testing of Hypothesis

Generally we want to test :

H_0 : X and Y are independent
Vs.
 H_1 : X and Y are not independent

Quite naturally : Reject H_0 if $|T|$ is large.

i.e. $|T| < C_\alpha$ where C_α
is the critical value
for level of significance
 $= \alpha$

NPTEL

So, generally, we want to test if X and Y are independent versus X and Y are not independent. So, what do we try to do? We try to compute the value of the T and we reject H_0 if T is large. That means the sample observation need not give us the exact value 0, but the maser of T should be less than some critical value C_α , where C_α is the critical value for level of sigdificance is equal to α .

(Refer Slide Time: 39:42)

Under H_0 , $\tau = 0$, since its distribution is symmetric.

Hence reject H_0 at level α if the observed value of T satisfies $|T| > t_{\alpha/2}$,
 where $P\{|T| \geq t_{\alpha/2} \mid H_0\} = \alpha$.

For small values of n the null distribution can be directly evaluated.

That is what we are saying under H_0 : $\tau = 0$ since the distribution is symmetric, hence reject H_0 at level α , if the observed value of tau satisfies the modulus of tau is greater than $t_{\alpha/2}$. For small values of n the null distribution, of course, can be directly evaluated. And the way we have done many problems through complete enumeration, we can also do that when the value of n is small.

(Refer Slide Time: 40:18)

Illustration 3

Consider $n = 3$.
 To make calculation simple we arrange X observations in sorted order. Hence there are 6 possible permutations for Y .

X	1	2	3	T
Y_1	1	2	3	1
Y_2	1	3	2	1/3
Y_3	2	1	3	1/3
Y_4	2	3	1	-1/3
Y_5	3	1	2	-1/3
Y_6	3	2	1	-1

changing $(x_1, y_1) (x_2, y_2) \dots (x_n, y_n)$
 x_i to i th $(r_1, y_1) (r_2, y_2) \dots (r_n, y_n)$
 rank
 2 if $r_1 < r_2 < \dots < r_n$
 $(1, y_1) (2, y_2) \dots (n, y_n)$

Now, let me give you another illustration. To make the calculation simple, we arrange the X observation in sorted order. So, we have arbitrary pairs (X_1, Y_1) , (X_2, Y_2) , up to (X_n, Y_n) . Suppose, we convert them to ranks, therefore, what we get rank of X_1, Y_1 , let me write it down for you changing X_i to its rank we get (r_1, y_1) , (r_2, y_2) , (r_n, y_n) and if r_1 is less

than r_2 less than r_3 less than r_n , then this gives us of the form $(1 y_1), (2 y_2), (n y_n)$ where 1 is the rank of X_1 , 2 is the rank of X_2 and n is the rank of X_n .

Or in other words, the calculation becomes simple if we arrange one observation, maybe the X in sorted order. Now, we have 3 pairs of observation let us sort X in increasing order 1, 2, 3 and then for Y we can have all the 6 possible permutations (1, 2, 3), (1, 3, 2), (2, 1, 3) etc and then we can compute the tau with respect to each permutation. And we will find that for this it is coming out to be 1 because they are perfectly concordant, for this order 3, 2, 1 there is perfect discordant and for others, we shall get the values like this.

(Refer Slide Time: 42:35)

Illustration 3

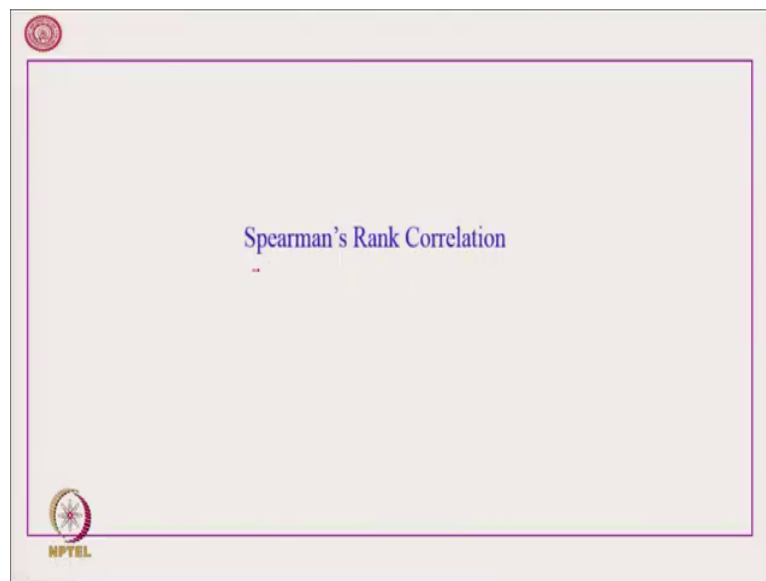
Consider $n = 3$.
To make calculation simple we arrange X observations in sorted order. Hence there are 6 possible permutations for Y .

X	1	2	3	T
Y_1	1	2	3	1
Y_2	1	3	2	-1/3
Y_3	2	1	3	1/3
Y_4	2	3	1	-1/3
Y_5	3	1	2	-1/3
Y_6	3	2	1	-1

Thus for relatively small values of n the Null Distribution of the Statistics τ can be evaluated

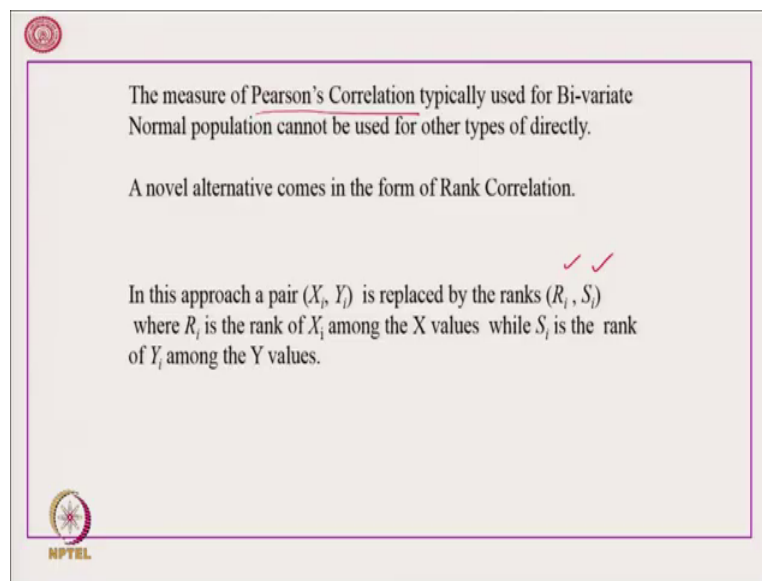
Thus, for relatively small values of n the null distribution of the statistic t can be easily evaluated because we are getting the probability distribution of t .

(Refer Slide Time: 42:55)



Let us now, consider the other metric that is Spearman's rank correlation, which also measures the association between 2 random variables which are jointly distributed.

(Refer Slide Time: 43:09)



So, the measure of Pearson's correlation typically used for bivariate normal population cannot be used for other types of data directly. That is what we started with that when we are computing the correlation, which is called Pearson correlation, it can be used for normal population. But, if the data is not of that type, then we can use a subtle variation of that, which is called rank correlation.

So, in this approach your pair (X_i, Y_i) is replaced by the ranks (R_i, S_i) , where R_i is the rank of X_i among the X observations and S_i is the rank of Y_i among the Y observations. We have already seen how we can replace a value with its rank, so we are not elaborating it much further.

(Refer Slide Time: 44:07)

 **Illustration**

For illustration five essays are being evaluated by two examiners as follows:

	E1	E2	E3	E4	E5
Examiner 1	95	93	70	68	65
Examiner 2	A	A+	B	B-	B+

Note:

1. The distributions are not Normal
2. Some values are not numeric but nominal/ordinal
3. The values do not match

 NPTEL

Suppose, we have the 5 essays graded by 2 different examiners, one has given numerical scores, others have given grades and we want to see to what extent the examiners agree with each other or in other words, the association between their evaluations. So, we can see that the distributions are not normal and there are values which are not numeric and also we can see that they are not having the similar opinion with respect to the 5 essays. Or in other words, there is a disagreement between the examiners.

(Refer Slide Time: 44:52)

Hence measuring association between the evaluation of the two examiners is not possible through Pearson Correlation which is computed as follows:

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$


Hence, the question of measuring their association is important. Here we invoke our earlier definition of $\rho_{X, Y}$, which we have already defined, so I am not going into that detail here. But for Spearman rank correlation the X_i will be replaced by R_i , and Y_i will be replaced by S_i .

(Refer Slide Time: 45:14)

	E1	E2	E3	E4	E5	
Examiner 1	1	2	3	4	5	← R
Examiner 2	2	1	4	5	3	← S

We rename the rank variables as R and S respectively and apply same correlation formula on R and S:



And therefore, from here we get the following. So, these are the ranks among the X's and these are the ranks among the Y's, which is the scores given by the examiner 2. Therefore, effectively by taking ranks we convert the data into numeric and this enables us to compute the correlation coefficient between R and S.

(Refer Slide Time: 45:44)

Hence the formula becomes:

$$RC = \frac{\text{Cov}(R, S)}{\sigma_R \sigma_S} = \frac{\sum_{i=1}^n (R_i - \bar{R}) * (S_i - \bar{S})}{\sqrt{\sum_{i=1}^n (R_i - \bar{R})^2} \sqrt{\sum_{i=1}^n (S_i - \bar{S})^2}}$$

The obvious disadvantage is a lot of information is lost.

But it has several advantages and easy to apply.

Which will now be of this formula and this is called rank correlation. It of course has certain disadvantage, because we are losing a lot of information, but it has certain advantages, because it is easy to apply.

(Refer Slide Time: 46:04)

1. $\bar{R} = \bar{S} = \frac{(n+1)}{2}$ ✓

Proof: Since both R and S takes values 1, 2, ...n

Why it is easy to apply, because of the simple thing that R takes values 1, 2, 3 up to n therefore, mean of R value is going to be n plus 1 by 2. Similarly, the mean of S is also going to be (n + 1)/ 2.

(Refer Slide Time: 46:24)


$$\begin{aligned} 2. \text{Var}(R) &= \text{Var}(S) = \frac{n(n^2-1)}{12} \\ \text{Proof: } \text{Var}(R) &= \sum_{i=1}^n (R_i - \bar{R})^2 = \sum_{i=1}^n R_i^2 - n\bar{R}^2 \\ &= \frac{n(n+1)(2n+1)}{6} - n\left(\frac{n+1}{2}\right)^2 \\ &= \frac{n(n+1)}{2} \left(\frac{(2n+1)}{3} - \frac{(n+1)}{2} \right) \\ &= \frac{n(n+1)}{2} \left(\frac{4n+2-3n-3}{6} \right) \\ &= \frac{n(n^2-1)}{12} \end{aligned}$$


What about the variance, the variance is coming out to be

$$\frac{n(n^2-1)}{12}$$

because it is very simple, because variance of R is equal to sigma over Ri minus R bar whole square

also we can write it as sigma Ri squared minus n times R bar square

that is the alternative formula for variance. Now, if we simplify them, we find that the variance of R is equal to

$$\frac{n(n^2-1)}{12}$$

In a very similar way, variance of S is equal to

$$\frac{n(n^2-1)}{12}$$

(Refer Slide Time: 47:06)

From the above two we get the value of the Rank Correlation:

$$\begin{aligned}
 RC &= \frac{\sum_{i=1}^n (R_i - \bar{R}) * (S_i - \bar{S})}{\sqrt{\sum_{i=1}^n (R_i - \bar{R})^2} \sqrt{\sum_{i=1}^n (S_i - \bar{S})^2}} \\
 &= \frac{12 \sum_{i=1}^n (R_i - \bar{R})(S_i - \bar{S})}{n^3 - n} \\
 &= \frac{12 \left[\sum_{i=1}^n R_i S_i - \bar{R} \sum_{i=1}^n S_i - \bar{S} \sum_{i=1}^n R_i + n \bar{R} \bar{S} \right]}{n(n^2 - 1)}
 \end{aligned}$$

Therefore, when we put these values in the formula for rank correlation, what we get it, sigma Ri minus R bar into Si minus S bar summation over i is equal to 1 to n divided by their standard deviations. Now, because these 2 values are same, we get n cubed minus n and what we are getting eventually is this formula 12 upon n into n square minus 1 multiplied by, as we open up this product with the summation we get

$$\frac{\sum_{i=1}^n R_i S_i - \bar{R} \sum_{i=1}^n S_i - \bar{S} \sum_{i=1}^n R_i + n \bar{R} \bar{S}}{n(n^2 - 1)} =$$

(Refer Slide Time: 47:59)

$$\begin{aligned}
 &= \frac{12 \left[\sum_{i=1}^n R_i S_i - n \bar{R} \bar{S} \right]}{n(n^2 - 1)} \\
 &= \frac{12 \left[\sum_{i=1}^n R_i S_i - n(n+1)^2/4 \right]}{n(n^2 - 1)} \\
 &= \frac{12 \sum_{i=1}^n R_i S_i}{n(n^2 - 1)} - \frac{3(n+1)}{n-1}
 \end{aligned}$$

Which is coming out to be

$$\frac{12 \left[\sum_i R_i S_i - n \cdot \bar{R} \bar{S} \right]}{n(n^2 - 1)}$$

which can be further simplified to

$$\frac{12 \sum_{i=1}^n R_i S_i}{n(n^2 - 1)} - \frac{3(n+1)}{n-1}$$

(Refer Slide Time: 48:22)

$$= \frac{12 \left[\sum_i R_i S_i - n \cdot \bar{R} \bar{S} \right]}{n(n^2 - 1)}$$

$$= \frac{12 \left[\sum_{i=1}^n R_i S_i - n(n+1)^2/4 \right]}{n(n^2 - 1)}$$

$$= \frac{12 \sum_{i=1}^n R_i S_i}{n(n^2 - 1)} - \frac{3(n+1)}{n-1}$$

Illustration:

	E1	E2	E3	E4	E5
Ex 1	1	2	3	4	5
Ex 2	2	1	4	5	3

Therefore RC =

$$\frac{12(2 + 2 + 12 + 20 + 15)}{5 \cdot (25 - 1)} - \frac{3 \cdot 6}{4}$$

= 5.1 - 4.5 = 0.6

Therefore, consider the problem that we are dealing with. We have the ranks of R's like this and the rank of S like this, therefore the rank correlation is coming out to be

$$\frac{12(2 + 2 + 12 + 20 + 15)}{5 \cdot (25 - 1)} - \frac{3 \cdot 6}{4}$$

which is coming out to be 0.6. So, that is the rank correlation or Spearman's rank correlation between these two pairs of observation.

(Refer Slide Time: 49:14)

An Alternative Formula

In this formula we consider the differences $D_i = R_i - S_i$ as follows:

First observe: D_i can be written as $D_i = (R_i - \bar{R}) - (S_i - \bar{S})$

Therefore

$$\begin{aligned} \sum_i D_i^2 &= \sum_i (R_i - \bar{R})^2 + \sum_i (S_i - \bar{S})^2 - 2 \sum_i (R_i - \bar{R})(S_i - \bar{S}) \\ &= \sum_i (R_i - \bar{R})^2 + \sum_i (S_i - \bar{S})^2 - 2 \sum_i (R_i - \bar{R})(S_i - \bar{S}) \\ &= 2 * \frac{n(n^2 - 1)}{12} - 2 \sum_i (R_i - \bar{R})(S_i - \bar{S}) \end{aligned}$$

Now, there is a alternative formula for that one. This is computed using the value of D_i , which is the difference in ranks for R_i and S_i . So, D_i we can write it as

$$(R_i - \bar{R}) - (S_i - \bar{S})$$

because, $\bar{R}$ is same as $\bar{S}$ by both are $(n+1)/2$. Therefore, summation of D_i square is equal to summation of square of these quantities, which we can open up and we get

$$\sum_i (R_i - \bar{R})^2 + \sum_i (S_i - \bar{S})^2 - 2 \sum_i (R_i - \bar{R})(S_i - \bar{S})$$

Now, this is giving you the notion of the variance and 2 of them that is how we get

$$2 * \frac{n(n^2 - 1)}{12} - 2 \sum_i (R_i - \bar{R})(S_i - \bar{S})$$

(Refer Slide Time: 50:15)



$$\Rightarrow 2 \sum_i (R_i - \bar{R})(S_i - \bar{S}) = 2 * \frac{n(n^2 - 1)}{12} - \sum_i D_i^2$$

$$\Rightarrow 12 \sum_i (R_i - \bar{R})(S_i - \bar{S}) = n(n^2 - 1) - 6 \sum_i D_i^2$$

Since we have already seen that

$$RC = \frac{12 \sum_{i=1}^n (R_i - \bar{R})(S_i - \bar{S})}{n^3 - n}$$

Together we get

$$RC = 1 - \frac{6 \sum_i D_i^2}{n(n^2 - 1)}$$


Therefore, $2 \sum (R_i - \bar{R})(S_i - \bar{S})$ we can write it as $2 * \frac{n(n^2 - 1)}{12} - \sum D_i^2$ or if we multiply this by 6 we get $12 \sum (R_i - \bar{R})(S_i - \bar{S})$ is equal to $n(n^2 - 1) - 6 \sum D_i^2$. Since, we have already seen that the rank correlation is equal to $12 \sum (R_i - \bar{R})(S_i - \bar{S})$ divided by $n^3 - n$. Therefore, together we can write that the rank correlation is equal to

$$1 - \frac{6 \sum_i D_i^2}{n(n^2 - 1)}$$

(Refer Slide Time: 51:12)



For illustration we take the same example:

	E1	E2	E3	E4	E5
Ex 1	1	2	3	4	5
Ex 2	2	1	4	5	3

Therefore:

D_i	-1	1	-1	-1	+2
-------	----	---	----	----	----

Hence

$$RC = 1 - \frac{6 \sum_i D_i^2}{n(n^2 - 1)} = 1 - \frac{6 * (1 + 1 + 1 + 1 + 4)}{120} = 1 - 0.4 = 0.6$$


Therefore, if we apply the same formula for the same data, then we get D_i is equal to -1, +1, -1, -1 and +2. Therefore, 1 minus 6 into sigma D_i squared is coming out to be

$$1 - 6 \times (1 + 1 + 1 + 1 + 4) / n \times n^2 - 1,$$

which is 5 into 24, that is 120 therefore, again we get the same value 0.6. So, that is how we calculate the Spearman's rank correlation for a given set of paired observations.

(Refer Slide Time: 51:59)



To test H_0 : X and Y are independent ✓
Vs.
 H_1 : X and Y are not independent, ✓

We reject the null hypothesis if the absolute value of rank coefficient RC is large i.e. reject H_0 if $|RC| \geq R_{\frac{\alpha}{2}}$ at level of significance α , where $R_{\frac{\alpha}{2}}$ is the upper tailed critical value corresponding to $\alpha/2$.



Again, we do the similar type of testing whether X and Y are independent or they are not independent, therefore we reject the null hypothesis. If the absolute value of the rank coefficient is somewhat large, that is it is greater than some threshold, otherwise we are going to accept that they are independent.

(Refer Slide Time: 52:22)

We now consider an example to illustrate how we can test the Null Hypothesis H_0 using the two measures of association.

Consider the following random sample of 6 pairs from a bivariate population of random variables X and Y.

(17.81, 20.48), (20.19, 18.13), (15.56, 18.50),
(22.55, 19.85), (21.50, 17.79), (19.00, 19.50)

Use Kendall's Tau and Spearman's Rank correlation coefficient to test if the two variables are independent against the general two-sided alternative of dependence.

We now consider an example to illustrate how we can test the null hypothesis H_0 using the two measures of association, that is Kendall's Tau and Spearman's rank correlation. So, consider the following random sample of 6 pairs from a bivariate population of random variables X and Y. So, the values are 17.81 and Y is equal to 20.48, second sample is X is 20.19 and Y is 18.13, like that this is the 6th sample X is 19 and Y is 19.50. So, the task is that to use Kendall's Tau and Spearman's rank correlation to test if the two variables are independent and alternative is that these two are not independent.

(Refer Slide Time: 53:36)

Kendall's Tau

We need to count the number of concordant pairs.

For that we arrange the pairs in increasing order of X:

X	15.56	17.81	19.00	20.19	21.50	22.55
Y	18.50	20.48	19.50	18.13	17.79	19.85

Now we count the number of times $Y_j - Y_i > 0$ for $1 \leq i < j \leq 6$ (in the sorted pairs)

Note that there are $\binom{6}{2} = 15$ pairs to be considered

The number of concordant pairs is $3 + 0 + 1 + 1 + 1 = 6$.

So, let us first discuss Kendall's Tau. We need to count the number of concordant pairs. So, for that, we first arrange the pairs in increasing order of X. So, you can see that 15.56, 17.81 like that up to 22.55 and Y's are kept accordingly as per the pairs to stop now, we count the

number of times $Y_j - Y_i$ is greater than 0, because we always know that if $j > i$, then $X_j > X_i$. So, we are looking at $Y_j > Y_i$ for $1 \leq i < j \leq 6$.

Now, there are 6C_2 , that is 15 pairs to be considered and we can see that the number of concurrent pairs is 6. How? So, we look at for the first one Y is 18.50 there are 1, 2 and 3 of them are bigger than this value, therefore, corresponding to the first pair, we get the value 3. For the second pair, it is 20.48, the value of Y , but none of the other one is greater than that, therefore, we get 0. In a similar way, we get 1 for this one, namely from here, 1 for 18.13, and 1 for 17.79. So, the number of concordant pairs is 6.

(Refer Slide Time: 55:39)

Since, there are no ties in the Y values, thus number of discordant pairs is : $15 - 6 = 9$.

Therefore observed value of T-statistic is: $P_c - P_d =$

$$\frac{6 - 9}{\binom{6}{2}} = -\frac{3}{15} = -0.20$$

Therefore, observed value of $|T| = 0.2$

This value now we need to check with tabulated Critical Value.

Here is a glimpse of the Table taken from:

<https://www.york.ac.uk/depts/maths/tables/kendall.pdf> ✓

Since there are no ties in the Y values, the number of discordant pair is going to be 15 minus 6, that is 9. Therefore, the observed value of T statistic, which is probability of concordance minus probability of discordance, which is coming out to be 6 minus 9 divided by 6C_2 , that is 15 and this gives us - 0.20. Therefore, of sound value of the T, which is the absolute value of the P_c minus P_d is coming out to be 0.2. So, that is the statistic that we have to use for acceptance or rejection of the null hypothesis. Therefore, what we do, we check it with the tabulated critical value. So, we have taken the table from this side, I have given it for your benefit, and we get the following.

(Refer Slide Time: 56:44)

n	Nominal α					
	0.10	0.05	0.025	0.01	0.005	0.001
4	1.000	1.000	.	.	.	.
5	0.800	0.800	1.000	1.000	.	.
6	0.600	0.733	0.867	0.867	1.000	.
7	0.524	0.619	0.714	0.810	0.905	1.000
8	0.429	0.571	0.643	0.714	0.786	0.857
9	0.389	0.500	0.556	0.667	0.722	0.833
10	0.378	0.467	0.511	0.600	0.644	0.778
11	0.345	0.418	0.491	0.564	0.600	0.709
12	0.303	0.394	0.455	0.545	0.576	0.667
13	0.308	0.359	0.436	0.513	0.564	0.641
14	0.275	0.363	0.407	0.473	0.516	0.604
15	0.276	0.333	0.390	0.467	0.505	0.581
16	0.250	0.317	0.383	0.433	0.483	0.567
17	0.250	0.309	0.368	0.426	0.471	0.544
18	0.242	0.294	0.346	0.412	0.451	0.529
19	0.228	0.287	0.333	0.392	0.439	0.509
20	0.221	0.274	0.326	0.379	0.421	0.495

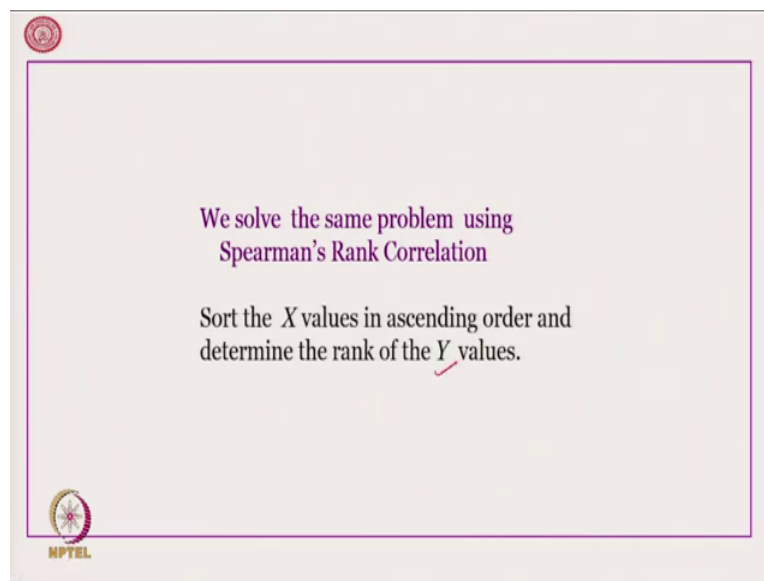
Note that here $n = 6$.
Therefore, for $\alpha = 0.05$,
i.e. $\frac{\alpha}{2} = 0.025$, the upper
tailed critical value for
0.025 is 0.867
and
for $\alpha = 0.01$, i.e. $\frac{\alpha}{2} = 0.005$
the upper tailed critical
value for 0.005 is 1.000

Since $0.20 < 0.867$
we do not reject H_0 for
 $\alpha = 0.05$ and $\alpha = 0.01$.

So, this is the sample from the table where corresponding to different n 's, I am showing some 4 to 20 and different values of α 0.1, 0.05, 0.025 etc. up to 0.001 we have been given the critical value. So, for our example, n is equal to 6, therefore, when α is equal to 0.05, α by 2 is 0.025. Therefore, the critical value that we can see is 0.867. Again, if we look at 1% level of significance, then α is equal to 0.01, therefore α by 2 is equal to 0.005 and the corresponding value given here is 1.

Now, our test statistic is 0.2 which is less than 0.867 and therefore naturally < 1 , therefore, we cannot reject the null hypothesis for α is equal to 0.05 and α is equal to 0.01.

(Refer Slide Time: 58:09)



Now, let us solve the same problem using Spearman's rank correlation. So, we have already sorted the X values in ascending order and let us determine the rank for the Y values.

(Refer Slide Time: 58:29)

Spearman's Rank						
X	15.56	17.81	19.00	20.19	21.50	22.55 ✓
R_i	1 ✓	2 ✓	3 ✓	4 ✓	5 ✓	6 ✓
Y	18.50	20.48	19.50	18.13	17.79	19.85
S_i	3 ✓	6 ✓	4 ✓	2 ✓	1 ✓	5 ✓
$D_i = R_i - S_i$	-2 ✓	-4 ✓	-1 ✓	2 ✓	4 ✓	1 ✓

Therefore, $\sum_{i=1}^6 D_i^2 = 42$ ✓

Hence $R = 1 - 6 \left[\frac{\sum_{i=1}^6 D_i^2}{6(6^2 - 1)} \right] = 1 - 6 \left[\frac{42}{6 \times 35} \right] = 1 - \frac{42}{35} = -0.2$

Therefore, Test Statistic = $|R| = 0.2$

So, let us focus on this table X is in increasing order, therefore, the rank of the elements are $R_1 = 1$ or $R_2 = 2$ up to $R_6 = 6$. Now, these are the corresponding Y values and let us see what are their ranks. So, this is rank 1 that is S_5 is equal to 1, S_4 is equal to 2, S_1 is equal to 3, S_3 is equal to 4 and S_6 is equal to 5 and S_2 is equal to 6. Therefore, what we get $R_i - S_i$ is equal to -2, -4, -1, 2, 4 and 1.

Therefore, as we know we compute sigma D_i square which is coming out to be 42 and therefore, the value of R is equal to

$$1 - (6 \times 42 / 6 \times 35)$$

that is

$$1 - 42 / 35 = -0.2.$$

Therefore, the test statistic R is also 0.2, because we take the absolute value of this.

(Refer Slide Time: 59:59)

Refer to the table of Critical values taken from:
<https://www.york.ac.uk/depts/maths/tables/spearman.pdf> ✓

n	Nominal α				
	0.10	0.05	0.025	0.01	0.005
4	1.000	1.000	-	-	-
5	0.800	0.900	1.000	-	-
6	0.657	0.829	0.886	0.943	1.000
7	0.571	0.714	0.786	0.893	0.929
8	0.524	0.643	0.738	0.833	0.881
9	0.483	0.600	0.700	0.783	0.833
10	0.455	0.564	0.648	0.745	0.794
11	0.427	0.536	0.618	0.709	0.755
12	0.406	0.503	0.587	0.678	0.727
13	0.385	0.484	0.560	0.648	0.703
14	0.367	0.464	0.538	0.626	0.679
15	0.354	0.446	0.521	0.604	0.654
16	0.341	0.429	0.503	0.582	0.635
17	0.328	0.414	0.488	0.566	0.618
18	0.317	0.401	0.472	0.550	0.600
19	0.309	0.391	0.460	0.535	0.584
20	0.299	0.380	0.447	0.522	0.570

Note that here $n = 6$.
 Therefore, for $\alpha = 0.05$,
 i.e. $\frac{\alpha}{2} = 0.025$, the upper
 tailed critical value for
 0.025 is 0.886
 and
 for $\alpha = 0.01$, i.e. $\frac{\alpha}{2} = 0.005$,
 the upper tailed critical
 value for 0.005 is 1.000
 Since $0.20 < 0.886$
 we do not reject H_0 for
 $\alpha = 0.05$ and $\alpha = 0.01$. ✓

Again as we before we look at the corresponding table, again I have given you the link for the table. And as before we look at for 0.025 and 0.005, the values that we can see corresponding to n is equal to 6 are 0.886 and 1. And therefore, since the observed value is less than 0.886, we cannot reject the H_0 for $\alpha = 0.05$ and $\alpha = 0.01$. So, that is how we use the Kendall's Tau or Spearman's rank correlation to test a hypothesis about whether the sample pairs are showing independence or not.

(Refer Slide Time: 61:03)

Large Number Approximations

Large number approximations.

(Refer Slide Time: 61:06)

Kendall's Tau

For large values of n (> 8) under the Null Hypothesis the Kendall's Tau coefficient may be approximated by Normal distribution as follows:

The random variable $Z = \frac{3\sqrt{n(n-1)}T}{\sqrt{2(2n+5)}}$ may be considered to be distributed as Standard Normal.

The Null Hypothesis may be accepted or rejected accordingly.



As usual, when n is slightly large, we can try to approximate that using standard normal distribution, with respect to Kendall's tau, what we find that if t is the observed value of the statistic, then

$$\frac{3\sqrt{n(n-1)}}{\sqrt{2(2n+5)}}$$

this whole multiplied by T may be considered as a standard normal distribution. We are not going to derive this thing, let us accept that and that null hypothesis will be accepted or rejected by using Z test with this value of Z that is obtained by a transformation of the value that we got from the sample for the statistic T .

(Refer Slide Time: 62:09)

Spearman Rank Coefficient

Under the Null hypothesis the random variable

$$Z = \frac{R\sqrt{n-1}}{\sqrt{1-R^2}}$$

has approximately Standard Normal Distribution, where R is the Spearman's Coefficient.

Also, the statistic $t = \frac{(R\sqrt{n-2})}{\sqrt{1-R^2}}$ approximately follows Student's t distribution with $n-2$ degrees of freedom.

It has been found to give more accurate results for moderate n .



With respect to Spearman rank coefficient, under the null hypothesis, the random variable Z is equal to R times root over n minus 1 has approximately a standard normal distribution, where R is the obtained value of the Spearman rank coefficient. However, it has been found that another statistic

$$(R\sqrt{n-2})/(1-R^2)$$

approximately follows Student's t distribution with $n - 2$ degrees of freedom.

And it has been found that for moderate values of n , that is n is not too large, then this T statistic gives a better result than the normal approximation. Okay friends, I stop here today. In the next class, I shall continue with some more discussions about different nonparametric tests. Thank You.