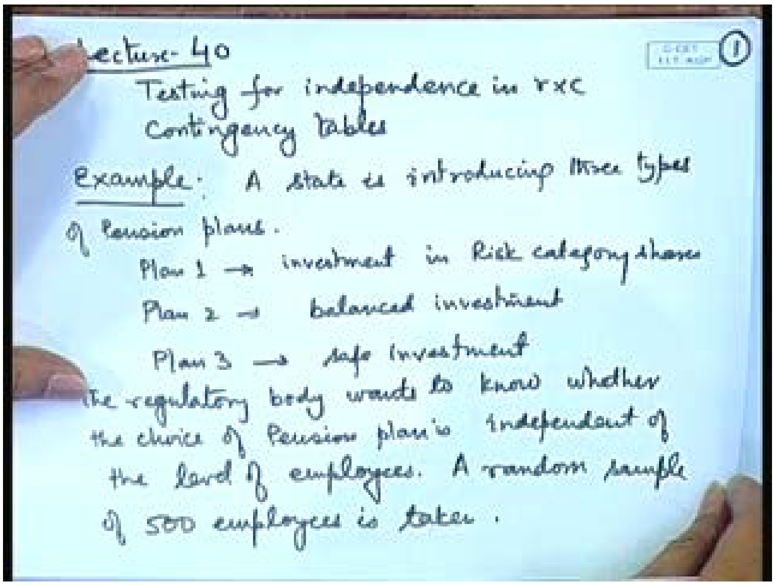
Probability and Statistics
Prof. Dr. Somesh Kumar
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture No. #40
Testing of Hypothesis-VIII

Today, we have discussed that we can do testing of hypothesis problems for situations such as testing for goodness of it that means, when a data set is given to us we want to check from which particular distribution it has come from. So, for this situation we have given a chi square test for goodness of it. Similarly, there is another situation where we have the data of the type which is categorical and we want to test whether the categories are independent.

(Refer Slide Time: 00:56)



ecture- 40 ①

Testing for independence in r×c contingency tables

Example: A state is introducing three types of Pension plans.

Plan 1 → investment in Risk category shares

Plan 2 → balanced investment

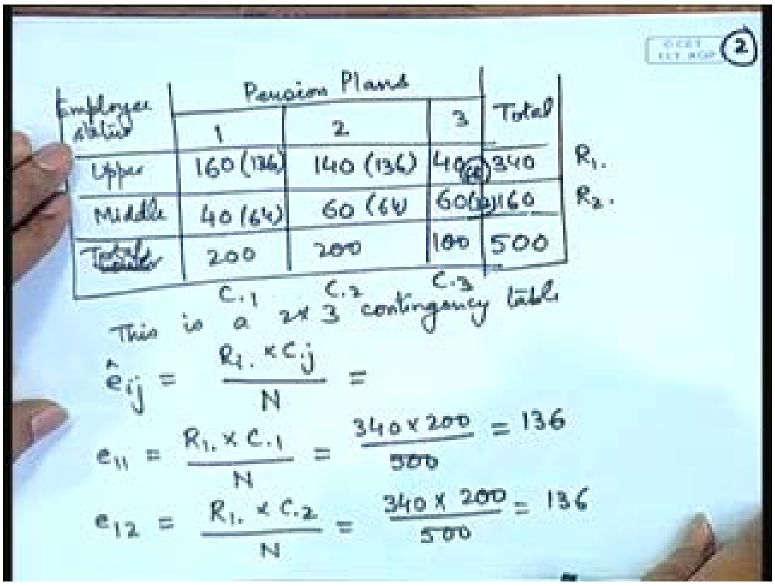
Plan 3 → safe investment

The regulatory body wants to know whether the choice of Pension plan is independent of the level of employees. A random sample of 500 employees is taken.

So, this topic we called as Testing for independence- testing for independence in r by c contingency tables. So, yesterday, I mentioned that we may have r categories distributed in the rows and c categories which are represented in the columns. We have observed frequencies Oij corresponding to i j cell and on the basis of this, we want to test whether the two categories are independent. Let me explain this through an example here. So, the problem is posed as follows: so, a state is introducing three types of pension plans. So, in the plan 1 the investment of the pension fund will be in risk category shares, in plan 2 this is balanced investment and plan 3 is say, for safe investment.

Now, whether the employees' preferences are affected by their hierarchical structure in the organization, that is what we have to check. So, the regulatory body wants to know whether the choice of pension plan is independent of the level of employees. So, a random sample of 500 employees is taken and we observe the following data.

(Refer Slide Time: 04:01)



Employee status	Pension Plans			Total	
	1	2	3		
Upper	160 (136)	140 (136)	40 (68)	340	R1.
Middle	40 (64)	60 (64)	60 (32)	160	R2.
Total	200	200	100	500	
	C.1	C.2	C.3		

This is a 2× 3 contingency table

$$\hat{e}_{ij} = \frac{R_{i.} \times C_{.j}}{N} =$$

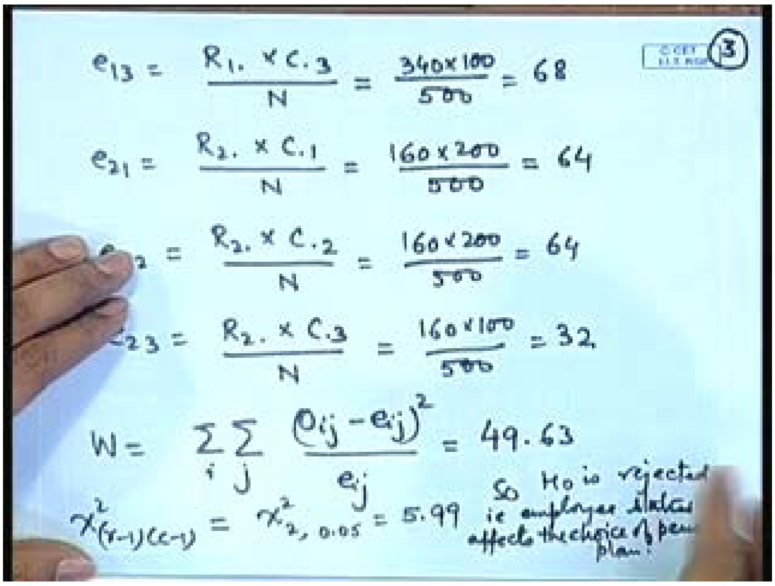
$$e_{11} = \frac{R_{1.} \times C_{.1}}{N} = \frac{340 \times 200}{500} = 136$$

$$e_{12} = \frac{R_{1.} \times C_{.2}}{N} = \frac{340 \times 200}{500} = 136$$

So, the choice of pension plans, we represent in the columns, that is plan 1, plan 2 and plan 3, and here we give employee status say, so, we have upper level, middle level and say, lower level. For the time being let me concentrate only, suppose upper level and middle level here, and the data of 500 is distributed as- 160 upper level employees give preference to pension plan 1, 140 give to pension plan 2 and 40 give to pension plan 3; in the middle level 40 give preference to pension plan 1, 60 to 2 and 60 to 3. So, if we calculate the row and column totals, they turn out to be 200, 200 and 100 and here it is 340, 160.

Now, we want to test whether the choice of pension plan is independent of the employee status. So, for this, this is a 2 by 3 contingency table, this is a 2 by 3 contingency table. So, the values of the eijs, they are calculated by Ri into C dot j divided by N. Now, here you see the row totals, that is R1 dot, R2 dot, this is C dot 1, C dot 2, C dot 3, they are given to us, so easily we can calculate say e11. So, e11 will be R1 dot into C dot 1 by N, now, in this particular case, it is 340 into 200 divided by 500, this value turns out to be 136- so, we can write these values here. Similarly, if I want to calculate e12, e12 is R1 dot into C dot 2 divided by N, that is equal to 340 into 200 divided by 500, that is again, 136.

(Refer Slide Time: 07:14)



$$e_{13} = \frac{R_{1.} \times C_{.3}}{N} = \frac{340 \times 100}{500} = 68$$

$$e_{21} = \frac{R_{2.} \times C_{.1}}{N} = \frac{160 \times 200}{500} = 64$$

$$e_{22} = \frac{R_{2.} \times C_{.2}}{N} = \frac{160 \times 200}{500} = 64$$

$$e_{23} = \frac{R_{2.} \times C_{.3}}{N} = \frac{160 \times 100}{500} = 32$$

$$W = \sum_i \sum_j \frac{(O_{ij} - e_{ij})^2}{e_{ij}} = 49.63$$

$$\chi^2_{(r-1)(c-1)} = \chi^2_{2,\,0.05} = 5.99$$

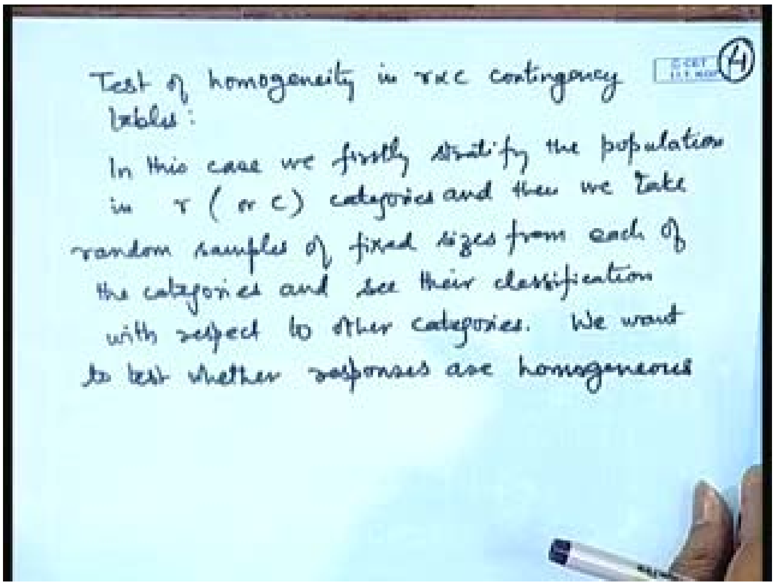
So H_0 is rejected
ie employee status
affects the choice of pen
plan!

Similarly, we can calculate e13 that is R1 dot into C dot 3 divided by N, that is equal to 340 into 100 divided by 500, that is equal to 68- so, this value is 68. (Refer Slide Time: 04:01) In a similar way, we can calculate the value corresponding to e21. So, e21 will be R2 dot into C dot 1 divided by N, that is equal to 160 into 200 divided by 500, that is equal to 64; e22, that will be equal to R2 dot into C dot 2 by N, that is equal to 160 into 200 divided by 500, that is again, 64; and e23, that is equal to R2 dot into C dot 3 divided by N, that is equal to 160 into 100 divided by 500, that is equal to 32. So, we can complete this table of eijs here. (Refer Slide Time: 04:01) So, this is 64, this is 64, this is 32.

Now, our formula for the W is sigma of Oij minus eij square divided by eij- so, these differences can be calculated. (Refer Slide Time: 04:01) So, for example, the first term is 160 minus 1 36, that is 24 square divided by 136 again, similar term here, this will become 4 square by 136, here it is 28 square by 30 by 68 and like that. So, we can calculate these terms. The overall W turns out to be 49.63. Now, here the calculated value of the chi square statistic will be on R minus 1 C minus 1 degrees of freedom; now, here 2 rows are there and 3 columns are there, so, this becomes chi square 2, and we look at say, 0.05, etcetera, then this is giving the value 5.99, which is much smaller- actually, we can calculate at a very small level of significance and this value will still be larger, so, H naught is rejected. What is H naught? H naught is the hypothesis that the row categories and the column categories are independent, that is H naught is rejected that means, the employee status affects the choice of pension plan.

There is another application of the chi square test. In the contingency table we have seen that we took a total sample of size N and then we saw the actual classification in the R C categories. But sometimes we fix the strata for example, we may take a fixed number from upper income group, upper level employees, we may take a fixed sample size from middle strata so that means, basically, we are doing the stratifafication of the population and then we take the sample and we want to see whether the responses of the different strata are homogenous.

(Refer Slide Time: 11:29)



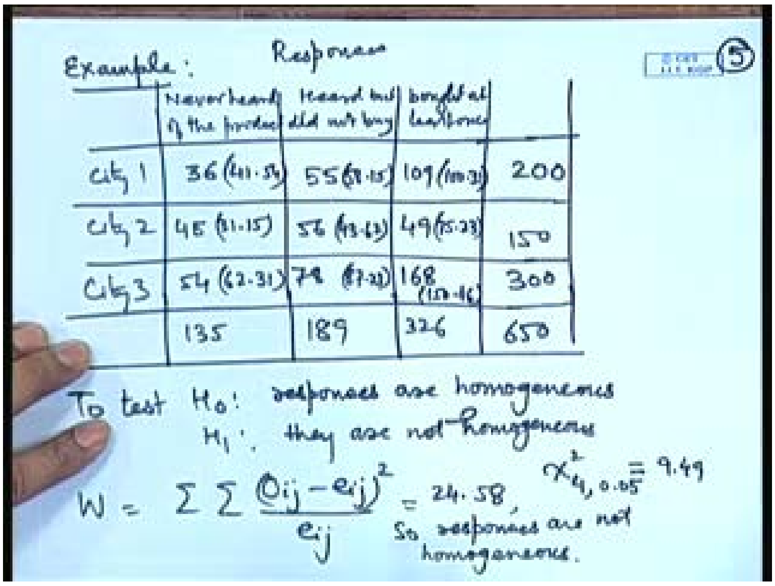
4

Test of homogeneity in $r \times c$ contingency tables:

In this case we firstly stratify the population in r (or c) categories and then we take random samples of fixed sizes from each of the categories and see their classification with respect to other categories. We want to test whether responses are homogeneous

So, in place of independence, this is termed as test of homogeneity in r by c contingency tables. So, in this case, we firstly, stratify the population in say, r or c categories and then we take samples of fixed sizes from each of the categories and see their classification with respect to other categories. We want to test whether responses are homogeneous. Now, you see, the sampling condition has been slightly modified, in the earlier case we took full sample size and that means, for the full population we take a sample, and then we see that to which i j cell they fall, now, here either the row sums or the column sums are fixed, and then we see that what is the frequency of each cell in each row or each column. So, the situation is slightly different, but the test of chi square goodness of fit, which we have given for independence, the same test is valid here also. Let me give this through an example.

(Refer Slide Time: 13:53)



Example:

	Responses			
	Never heard of the product	Heard but did not buy	bought at least once	
city 1	36 (41.54)	55 (58.15)	109 (100.31)	200
city 2	45 (31.15)	56 (43.63)	49 (75.23)	150
City 3	54 (62.31)	78 (87.23)	168 (150.46)	300
	135	189	326	650

To test H_0: responses are homogeneous
H_1: they are not homogeneous

$$W = \sum \sum \frac{(O_{ij} - e_{ij})^2}{e_{ij}} = 24.58, \quad \chi^2_{4, 0.05} = 9.49$$

So responses are not homogeneous.

So, a new product is introduced in the market and now we want to see whether it has the same level of effect in different towns of the country, or different regions of the state, or different areas of the city. So, in this particular case, we consider the response to the product in three different cities that means, the customer, or you can say the responder must have already purchased the product, or he might have heard about the product, but not purchased, or he might not have heard the product- so, the responses are based on a survey. Now, here what we do, we fix a number of respondents in each city, rather than we merge the data of the three cities and then taking a random sample, from each city we take a fixed sample size. So, let me present the data in the following form: so, we have city 1, city 2 and city 3, we fix that we are taking 200 respondents from city 1, we are taking 150 respondents from city 2 and we are taking 300 respondents from city 3.

This choice of the numbers may depend upon various factors, for example, the resources of the surveyor, the size of the city, for example, city 3 may be a much larger town compared to city 2 and city 1 may be somewhere in the middle as far as the population are concerned, or one may look at the consumption levels in different cities, based on that one may see such things- the total sample size from each strata may be decided on the basis of that. The responses are as follows: so, the respondent might never have heard of the product, heard, but did not buy or he might have bought at least once. So, the following observed data is there 36, 55, 109, 45, 56, 49, 54, 78 and 168.

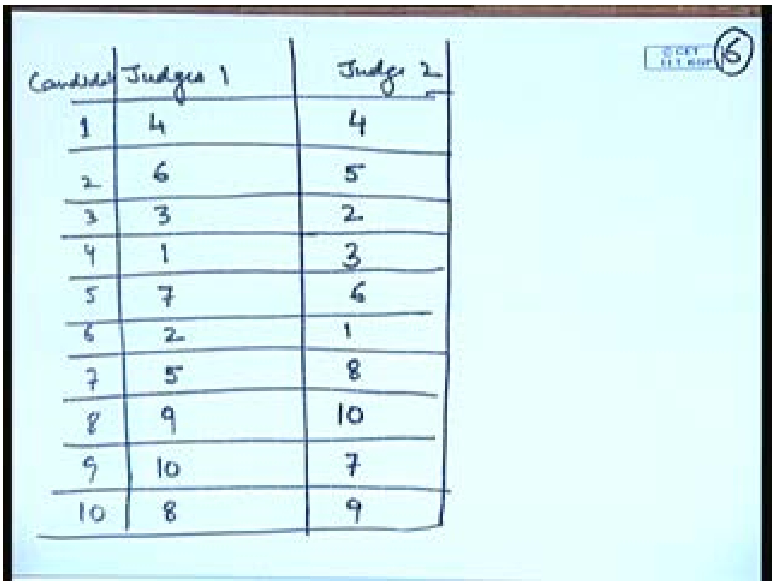
Now, to test whether the responses are homogenous against the hypothesis they are not homogeneous. We apply the same test, chi square test, that is double summation Oij minus eij square by eij. So, we will calculate the column sums, this is 135, 189, 326- the total sample size is 650. So, based on this we can calculate like, 200 into 135 divided by 165, that is say 41.54; 200 into 189 divided by 650, that is say 58.15 and so on; 100.31, 31.15, 43.62, 75.2 3, 62.31, 87.23, 150.4 6. So, based on these calculations of Oijs and eijs, we can evaluate this W and it turns out to be 24.58.

Now, if you look at chi square value, here the degrees of freedom will be 3 minus 1 into 3 minus 1, that is 4, suppose we take at 5 percent level of significance, this value is 9.4 9. Naturally, you can see that W is much larger than this. So, we conclude that H naught is rejected that means, responses are not homogenous. So, you can see here that this chi square test for goodness of fit is applicable in various situations, we are able to test for goodness of fit that means, whether the given data fits a given distribution, we can test for independence of the two types of classifications in a contingency table, we may test for the homogeneity of the responses in a contingency table situation- so, there are various applications.

Let me, we will come back to this again; firstly, let me introduce to further measures of correlation. Here, we have seen the Karl Pearson measure of correlation, which is calculated as covariance divided by the standard deviations of the two variables. Now, this measure of Karl Pearson, it is completely dependent upon the numerical observations of the variables concerned. For example, we may be looking at the relationship between the heights of say, parents with the heights of the children; we may be concerned with the expenditure on the health care by families corresponding to their per capita income, etcetera.

So, here actual measurements are required. But there are many situations in real life where the numerical values of the data are not very important, we may be simply concerned with say, ranks of the values, or we may be concerned about the increasing or decreasing trend of the values.

(Refer Slide Time: 21:30)

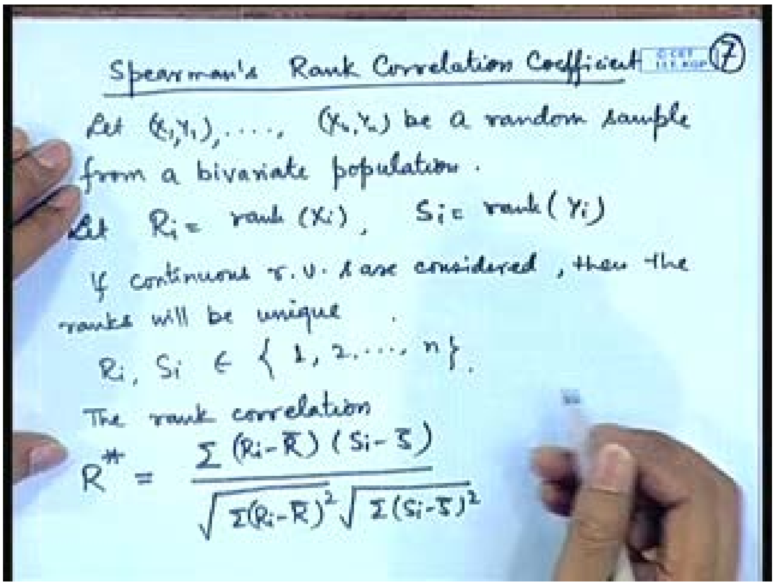


Candidate	Judges 1	Judge 2
1	4	4
2	6	5
3	3	2
4	1	3
5	7	6
6	2	1
7	5	8
8	9	10
9	10	7
10	8	9

For example, two judges give ranks to a set of participants in a certain competition. So, now, they are not telling that for example, it could be a selection procedure. So, in a selection procedure suppose 10 candidates are there and there are two judges- so, judge 1 and judge 2. Now, we have say candidates here say 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. So, rather than giving the scores their ranks are mentioned for example, judge 1, he ranks the candidate number say, 4 as 1 whereas, judge 2 ranks say, candidate number say, 6 as 1 likewise, they give ranks to all the candidates, the candidate number 6 may be ranked 2 here, the candidate number 3 may be ranked 3, the candidate number 3 may be ranked 2 by this, and this may be ranked 3, the candidate number 1 may be ranked 4 by say, both of them, the candidate number 2 may be ranked 6th here, and may be 5th here, candidate number 7th may be 5th here, 6th here, this may be 7th here say, this may be 7th here, may be this is 8th, this could be 8th here, this could be 9th here, may be this is 9, this is 10th , this is 10th .

Now, an impartial observer would like to see whether the selection has been fair or not that means, whether the two judges are working in concordance or discordance. So, here in place of the numerical values for example, the judges might have given some marks before giving the ranks to these candidates, but those are not important because what is important is the selection procedure is unbiased or not. So, in that case it is beneficial to look at the ranks and the correlation between these ranks. So, there is a procedure, or you can say term called Spearman’s rank correlation coefficient, which is introduced for this purpose.

(Refer Slide Time: 23:55)



Spearman's Rank Correlation Coefficient (7)

Let $(X_1,Y_1), \ldots, (X_n,Y_n)$ be a random sample from a bivariate population.

Let $R_i =$ rank (X_i), $S_i =$ rank (Y_i)

If continuous r.v.s are considered, then the ranks will be unique.

$R_i, S_i \in \{1, 2, \ldots, n\}$.

The rank correlation

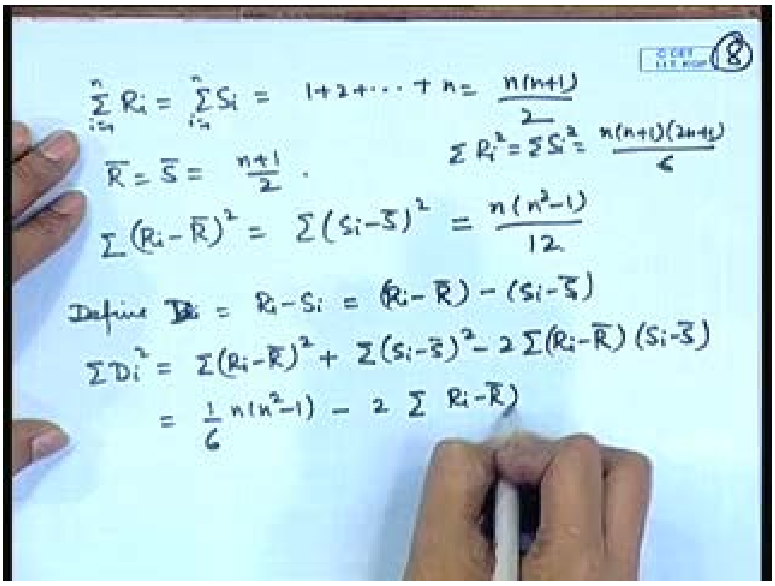
$$R^* = \frac{\Sigma (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\Sigma (R_i - \bar{R})^2} \sqrt{\Sigma (S_i - \bar{S})^2}}$$

Spearman’s Rank Correlation Coefficient. So, let X1 Y1, X2 Y2, XN YN be a random sample from a bivariate population. Let Ri be equal to rank of Xi and Si be say rank of Yi. So, what is the meaning of rank? We arrange this Xis and Yis in a increasing order. So, the smallest value is given rank 1, the second smallest is give rank 2, like that. So, now, the data may be discrete or continuous- if the data is continuous, then the probability of equality of two observations will be 0.

So, in that case the rankings will be unique that means, all the ranks from 1 to n will be allotted. So, if continuous random variables are considered for example, I considered heights, weights, incomes, so, these are all can be considered as continuous random variables, then the ranks will be unique. So, Ri, Si both will belong to the set 1, 2, n. If we want to calculate the correlation coefficient between ranks here, I will give a notation say R star because R we used for Karl Pearson correlation coefficient, so, this will be sigma Ri minus R bar into Si minus S bar divided by sigma Ri minus R bar square, square root sigma Si minus S bar square. So, basically what we are doing, in place of Xi and Yi we are using Ri and Si. So, from the same set of data, in place of using the numerical values, or numerical measurements corresponding to X and Y random variables now we are making use of their ranks.

Now, if we are making use of the ranks and if we are making this assumption that continuous random variables are there, then these values of R and S are between 1 to n only, and all the values are considered here. That means, if I look at sigma Ri, this will be the sum of the numbers 1, 2 up to n. Similarly, Sis, they will be 1 to n. Therefore, all these quantities can be actually evaluated.

(Refer Slide Time: 27:34)



8

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

$$\bar{R} = \bar{S} = \frac{n+1}{2}. \qquad \sum R_i^2 = \sum S_i^2 = \frac{n(n+1)(2n+1)}{6}$$

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

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

$$\sum D_i^2 = \sum (R_i - \bar{R})^2 + \sum (S_i - \bar{S})^2 - 2\sum (R_i - \bar{R})(S_i - \bar{S})$$

$$= \frac{1}{6} n(n^2-1) - 2 \sum R_i - \bar{R})$$

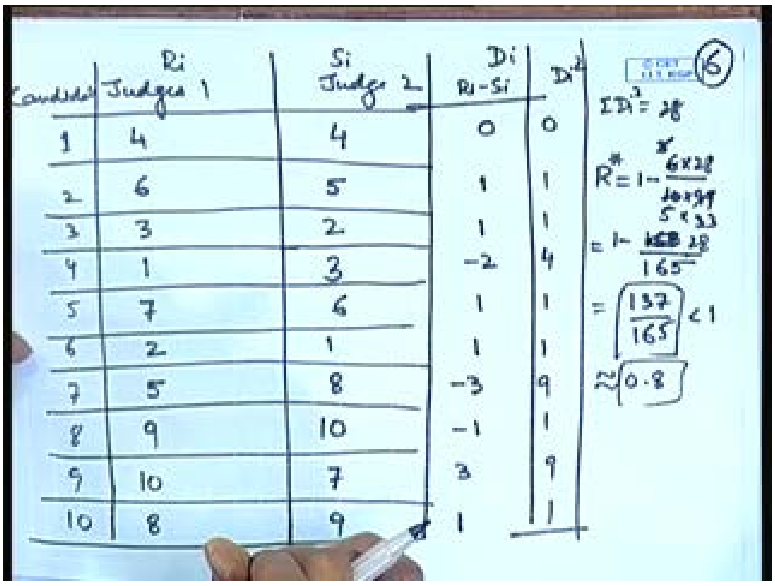
So, we can see here sigma Ri or sigma Si, i is equal to 1 to n, this is nothing but 1 plus 2 plus up to n, that is n into n plus 1 by 2. So, R bar and S bar they will become n plus 1 by 2. Similarly, we can calculate sigma Ri minus R bar whole square or sigma Si minus S bar whole square, because sigma Ri square or sigma Si square that will become n into n plus 1 into 2 n plus 1 by 6- that is the sum of squares of the first n integers. So, these values can be easily calculated and it turns out that this is n into n square minus 1 by 12.

So, if we substitute these values in this particular formula for R star, this turns out to be 12 into sigma Ri Si divided by n into n square minus 1 minus 3 into n plus 1 by n minus 1. Now, so, one thing is one can calculate sigma Ri Si, that is product of the ranks and substitute here. But there is a simplified procedure for this, which I will be developing now.

You define, now, say Di to be the differences in the, let me write capital Di differences in the ranks Ri minus Si, so, this of course, you can write as Ri minus R bar minus Si minus S bar- because R bar and S bar are same, so, they will cancel out. So, if I look at sigma of Di square, this turns out to be sigma Ri minus R bar square plus sigma Si minus

S bar whole square minus twice sigma Ri minus R bar into Si minus S bar. So, this turns out to be 1 by 6 n into n square minus 1 minus twice sigma Ri minus R bar Si minus S bar. So, from here sigma Ri Si can be substituted in terms of sigma Di square. So, if we substitute that here in this particular formula we get R star is equal to 1 minus 6 sigma Di square divided by n into n square minus 1. So, this particular term is called Spearman's correlation coefficient. So, the only difference is that we are not using the numerical values corresponding to the random variables, but the ranks of those values in the set of data that is given there.

(Refer Slide Time: 30:53)



Candidate	R_i Judges 1	S_i Judge 2	D_i $R_i - S_i$	D_i^2
1	4	4	0	0
2	6	5	1	1
3	3	2	1	1
4	1	3	-2	4
5	7	6	1	1
6	2	1	1	1
7	5	8	-3	9
8	9	10	-1	1
9	10	7	3	9
10	8	9	1	1

6

$\Sigma D_i^2 = 28$

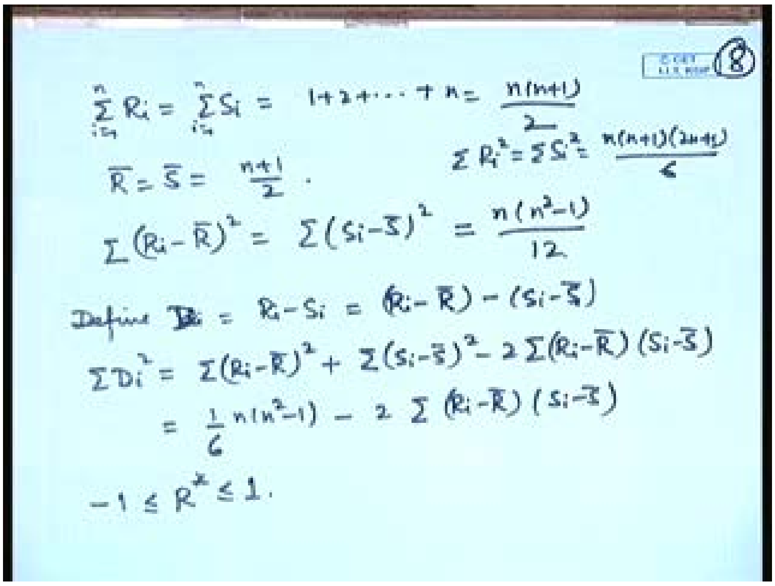
$$R^* = 1 - \frac{6 \times 28}{10 \times 99}$$
$$5 \times 33$$
$$= 1 - \frac{\cancel{168}\ 28}{165}$$
$$= \boxed{\frac{137}{165}} < 1$$
$$\approx \boxed{0.8}$$

So, as an example let me calculate in this particular case. So, here this can be considered as Ri and this can be considered as Di, Si. So, let us calculate here Dis, that is Ri minus Si. So, this is 0, this is 1, 1, minus 2, 1, 1, minus 3, minus 1, 3 and 1. So, if we look at Di square, that is 0, 1, 1, 4, 1, 1, 9, 1, 9, 1. So, sigma Di square is equal to 1 plus 1 plus 4, 28. So, now, if we substitute in this particular formula, that is 1 minus 6 sigma Di square divided by n into n square minus 1, so, the value of R star, that is equal to 1 minus 6 into 28 divided by 10 into 99, which can be simplified little bit, 1 minus- so, that is equal to, some value will come, 3, this is 5, this is 33- so, we get here 168 divided by 1, I am sorry, this is 28 divided by 165, that is equal to 137 by 165- so, which is of course, less than 1, but it is near about, if we calculate the actual value, it will be 0.8 approximately.

So, this is showing a high degree of correlation between the choices of, or you can say the preferences of judges 1 and 2 as far as the selection of the candidates is concerned.

That means, the candidates whom judge 1 gives a higher rank, the judge 2 also tends to give higher rank to the candidate, the candidates whom judge 1 gives lower ranks, the judge 2 also tends to give lower ranks to those candidates. So, there is a high degree of you can say, accordance, or concordance, or concurrence between the two judges here. Since this formula for the rank correlation is calculated from the Karl Pearson correlation coefficient therefore, the properties which the correlation coefficient satisfies also true here.

(Refer Slide Time: 34:18)



8

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

$$\bar{R} = \bar{S} = \frac{n+1}{2}. \qquad \sum R_i^2 = \sum S_i^2 = \frac{n(n+1)(2n+1)}{6}$$

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

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

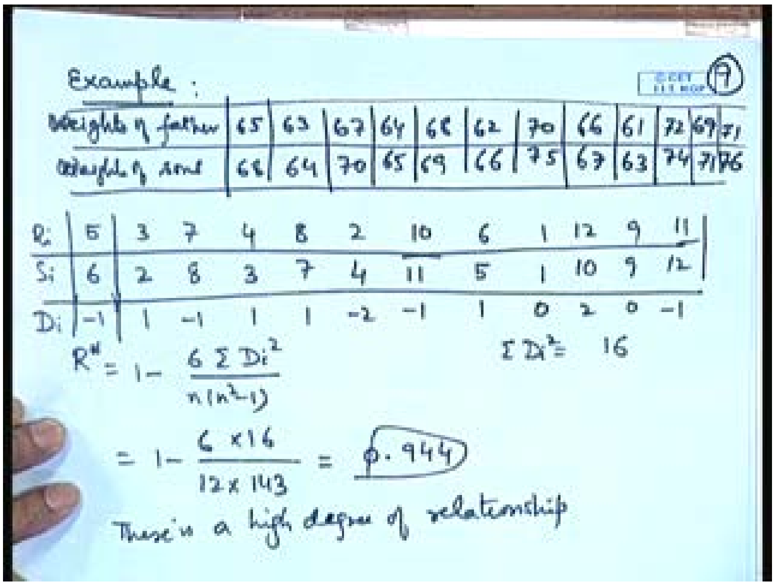
$$\sum D_i^2 = \sum (R_i - \bar{R})^2 + \sum (S_i - \bar{S})^2 - 2\sum (R_i - \bar{R})(S_i - \bar{S})$$

$$= \frac{1}{6} n(n^2-1) - 2\sum (R_i - \bar{R})(S_i - \bar{S})$$

$$-1 \le R^* \le 1.$$

That means, we will have in general, minus 1 less than or equal to R star less than or equal to 1, and this will be then again, considered as a measure of linear relationship, but this is relationship between the ranks. That means, for example, if we are considering R star is equal to 1 that means, this is showing a perfect positive linear relationship between the ranks R and S. Similarly, minus 1 will denote perfect negative linear relationship between the ranks R and S. So, all the observations related to the Karl Pearson correlation coefficient are valid here also, the only difference is that the correlations are calculated for the ranks rather than the raw values here.

(Refer Slide Time: 35:20)



Example :

7

Weights of father	65	63	67	64	68	62	70	66	61	72	69	71
Weights of sons	68	64	70	65	69	66	75	67	63	74	71	76

Ri	5	3	7	4	8	2	10	6	1	12	9	11
Si	6	2	8	3	7	4	11	5	1	10	9	12
Di	-1	1	-1	1	1	-2	-1	1	0	2	0	-1

$$R^{*} = 1 - \frac{6 \Sigma Di^2}{n(n^2-1)} \qquad \Sigma Di^2 = 16$$

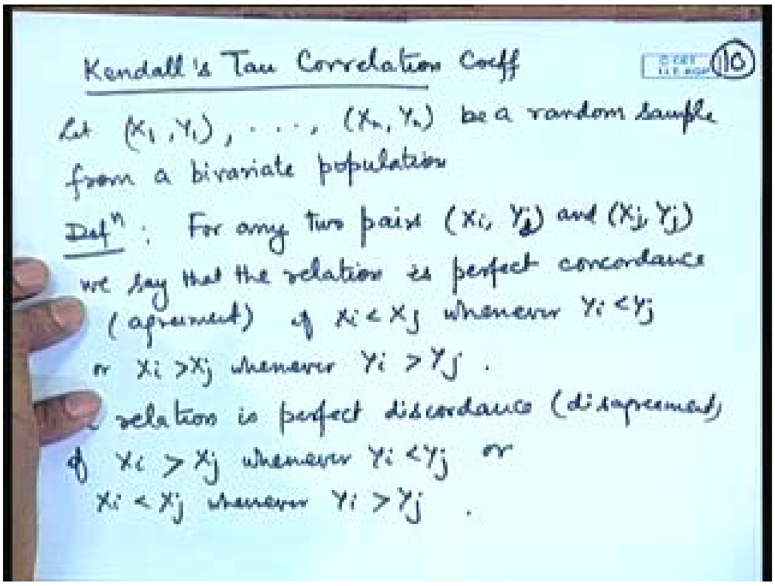
$$= 1 - \frac{6 \times 16}{12 \times 143} = \boxed{0.944}$$

There's a high degree of relationship

We will just consider one more example. So, we have heights of fathers and heights of say, sons and the data is as follows, sorry, this is not heights this is weights. So, there are 12 pairs 65-68, 63-64, 67-70, 64-65, 68-69, 62- 66 , 70-75, 66- 67, 61- 63, 72-74, 69-71, 71-76. Now, rather than calculating the correlation between the raw values of the weights we will consider their ranks, the reason is that we want to say that the heavier fathers have heavier sons and lighter fathers have lighter sons. So, if you want to say that, we just look at the ranks. So, if you look at the ranks Ris and Sis here, it turns out 5, so, from lightest to the heaviest we arrange 5, then this is 6, this is 3, this is 2, 7, 8, 4, 3, 8, 7, 2, 4, 10, 11, 6, 5, 1, 1 that is, the highest, the heaviest father has the heaviest son 12, 10, 9, 9, 11,12.

We want to see the concordance or not. So, if we look at the R star, further we need to calculate the differences, Di are minus 1, 1, minus 1, 1, 1, minus 2, minus 1, 1, 0, 2, 0 and minus 1. So, here if we calculate sigma Di square, that is equal to 16. So, 1 minus 6 into sigma Di square by n into n square minus 1, that is equal to 1 minus 6 into 16 divided by 12 into 143, after simplification this value is 0.944. So, you can see that there is a high degree of relationship, or you can say linear relationship in the ranks of the weights of the fathers and the sons. So, we can safely conclude that the heavier fathers have heavier sons.

(Refer Slide Time: 39:57)



Kendall's Tau Correlation Coeff 10

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from a bivariate population

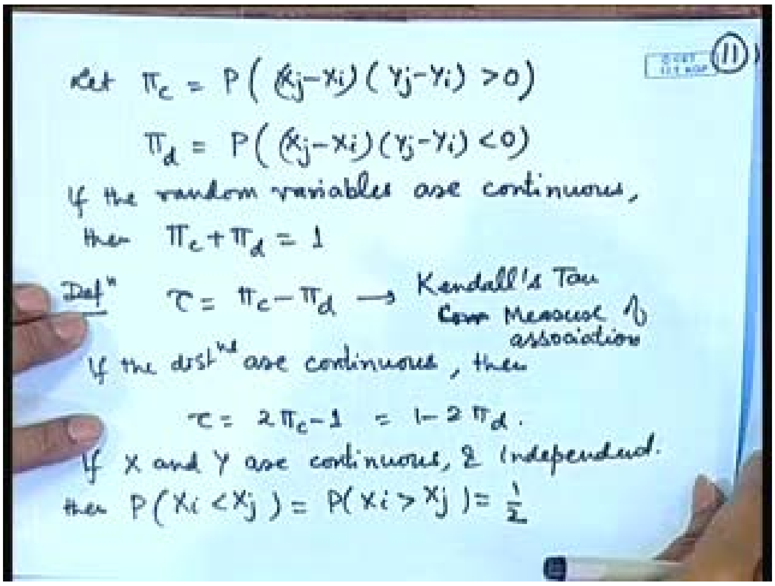
Def^n : For any two pair (X_i, Y_j) and (X_j, Y_j) we say that the relation is perfect concordance (agreement) if $X_i < X_j$ whenever $Y_i < Y_j$ or $X_i > X_j$ whenever $Y_i > Y_j$.

relation is perfect discordance (disagreement) if $X_i > X_j$ whenever $Y_i < Y_j$ or $X_i < X_j$ whenever $Y_i > Y_j$.

Yet another measure of correlation where in place of the raw values we look at only their magnitudes is Kendall’s tau correlation coefficient. In the ranks we looked at the position of the numerical value in the ordering of the observations, here in the Kendall’s tau we look at the side. That means, if we are having two values say Xi and Xj, then if Xj is on a higher side of Xi, then whether Yj is on the higher side of Yi or in the reverse. So, this is called concordance or accordance. So, let us write let X1 Y1, X2 Y2, XN YN be a random sample from a bivariate population. So, we have the following definition for concordance or discordance for any two pairs Xi Yj and Xj, sorry, Xi Yi and Xj Yj. We say that the relation is perfect concordance- so, concordance here means agreement, they are in the agreement- if Xi is less than Xj whenever Yi is less than Yj, or Xi is greater than Xj whenever Yi is greater than Yj. The relation is perfect discordance or disagreement if Xi is greater than Xj whenever Yi is less than Yj, or Xi is less than Xj whenever Yi is greater than Yj.

So, like the ranks this concordance or discordance also gives some indication of the kind of relationship that they are having. So, we create a measure of the concordance or discordance based on the following.

(Refer Slide Time: 42:12)



Let $\pi_c = P\left((X_j - X_i)(Y_j - Y_i) > 0 \right)$

$\pi_d = P\left((X_j - X_i)(Y_j - Y_i) < 0 \right)$

If the random variables are continuous, then $\pi_c + \pi_d = 1$

Def^n $\tau = \pi_c - \pi_d \longrightarrow$ Kendall's Tau Corr Measure of association

If the dist^{ns} are continuous, then

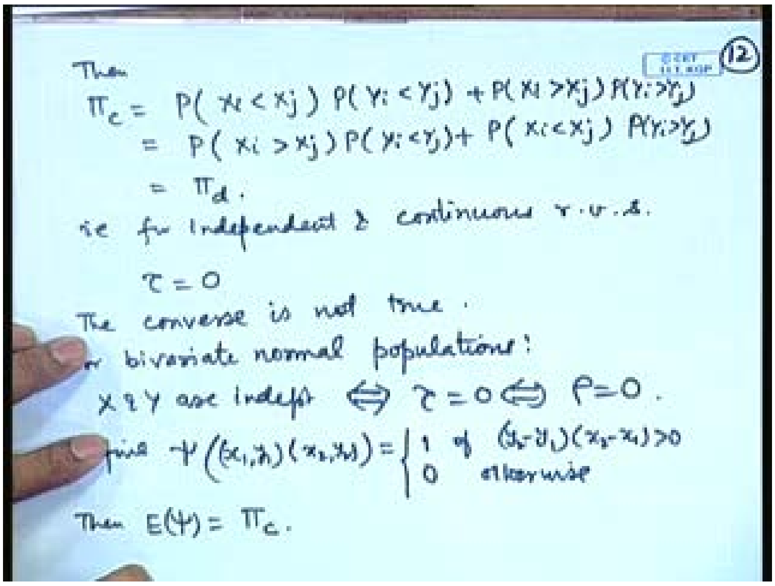
$\tau = 2\pi_c - 1 = 1 - 2\pi_d$.

If X and Y are continuous, & independent.

then $P(X_i < X_j) = P(X_i > X_j) = \frac{1}{2}$

Let us define pi C is equal to probability of Xj minus Xi into Yj minus Yi greater than 0- you can see this is the probability of concordance. And we can consider pi d as the probability of discordance of course, we may have to include equality at 1 place to exhaust all the possibilities. So, if the random variables are continuous, then you will have pi C plus pi d is equal to 1 because the probability of equal to 0 will be negligible or 0. So, we define tau, that is equal to pi C minus pi d- this is called Kendall’s tau correlation coefficient, or measure of association. If the distributions are continuous, then we may take tau is equal to twice pi C minus 1 by putting pi d is equal to 1 minus pi C, or you can say it as 1 minus twice pi d. Now, if X and Y are continuous and independent, then by symmetry you must have probability of Xi less than Xj is equal to probability of Xi greater than Xj, that is equal to half.

(Refer Slide Time: 44:25)



12

Then

$$\pi_c = P(x_i < x_j)\, P(Y_i < Y_j) + P(x_i > x_j)\, P(Y_i > Y_j)$$

$$= P(x_i > x_j)\, P(Y_i < Y_j) + P(x_i < x_j)\, P(Y_i > Y_j)$$

$$= \pi_d .$$

ie for independent & continuous r.v.s.

$\tau = 0$

The converse is not true.

or bivariate normal populations:

X & Y are indep $\Leftrightarrow \tau = 0 \Leftrightarrow \rho = 0$.

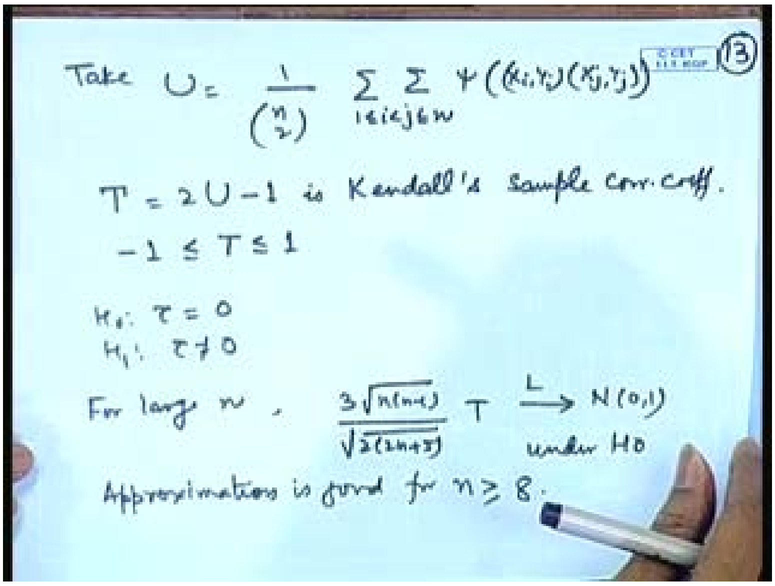
fine $\psi((x_1,y_1)(x_2,y_2)) = \begin{cases} 1 & \text{if } (y_2-y_1)(x_2-x_1)>0 \\ 0 & \text{otherwise} \end{cases}$

Then $E(\psi) = \pi_c$.

So, in that case, this pi C value will be equal to probability of Xi less than Xj into probability Yi less than Yj plus probability of Xi greater than Xj into probability of Yi greater than Yj, and because of the symmetric this can be then written as Xi greater than Xj into probability of Yi less than Yj, that is probability of Xi less than Xj into probability of Yi greater than Yj which is nothing but pi d. So, that is, for independent and continuous random variables tau is equal to 0- like in the Karl Pearson correlation coefficient, for independent random variables the Karl Pearson correlation coefficient was 0.

The converse of this is not true. However, as in the Karl Pearson case the bivariate normal distribution has a peculiar property that independence and the correlation is 0 or equivalent. The same thing is true here also. For bivariate normal populations X and Y are independent if and only if tau is 0, if and only if rho is 0. We look at a sample measure for, or you can say sample estimate for tau, so, we can define like this. Define a psi function for X1 Y1 and X2 Y2 pair as 1 if they are in concordance otherwise, you define it to be 0 then, naturally expectation of psi is equal to pi c.

(Refer Slide Time: 47:04)



13

Take $U = \frac{1}{\binom{n}{2}} \sum\sum_{1 \le i < j \le n} \psi((X_i, Y_i)(X_j, Y_j))$

$T = 2U - 1$ is Kendall's sample corr. coeff.

$-1 \le T \le 1$

$H_0: \tau = 0$

$H_1: \tau \ne 0$

For large n, $\frac{3\sqrt{n(n-1)}}{\sqrt{2(2n+5)}} T \xrightarrow{L} N(0,1)$ under H0

Approximations is good for $n \ge 8$.

So, we can consider an average of the all such pairs that means, we may consider U is equal to 1 by nC2 for all nC2 pairs of Xi Yi and Xj Yj for 1 less than or equal to i less than j less than or equal to n. Then, T, that is equal to 2U minus 1 is Kendall's sample correlation coefficient- again, we will have the value of T between minus 1 and 1. Of course, one may like to do a carry out a test of hypothesis whether this theoretical measure of association, tau, that is H naught, tau is equal to 0 against tau is not equal to 0. One can have a small sample test, one can work out the distribution of U, in fact, Kendall has tabulated the distribution of, the points of the distribution of T. However, a large sample test is given based on the approximation that the distribution of 3 into root n by n minus 1 root twice into 2n plus (( )) 5 into T is asymptotically normal 0, 1 under H naught- so, one can use this for testing the hypothesis tau is equal to 0, and it has been observed that approximation is good for n greater than or equal to 8.

Let me give an over view of the topics that we have covered in this particular course till now. We have, we started with the probability, the basic definition of the probability, we gave some basic methods such as the relative frequency definition, the classical definition. Then, we saw that a general framework based on the (()) axiomatic definition can be given for the probability. And we saw various tools, or you can say various results related to the basic probability like addition rule, multiplication rule, conditional probability byes theorem etcetera. Then, we introduced the concept of random variables, we saw the types of random variables that we have, discrete random variables, we have continuous random variables, we may have mixed random variables. Then, we talked about the probability distributions of random variables. The probability distributions of the random variables can be described by using probability mass functions, probability density functions and cumulative distribution functions. We also saw the characteristics such as mean, variance, third order moments, fourth order moments, the moment generating function, measures of symmetry, measures of peakedness, which are called measures of estimation kurtosis. We also saw some sort of probability bounds in the form of Chebyshev’s inequality, we discussed the jointly distributed random variables that means, in place of one if we have two random variables, or we have several random variables.

We can talk about the conditional distributions and the marginal distributions, the joint correlation coefficient, the joint moment generating function, etcetera. In particular, we studied special distributions, in the univariate case we studied discrete uniform distribution, binomial distribution, poison distribution, which arise in bernoullian trials, or in a sequence of events which can be described by a Poisson process. We looked at distribution such as hyper geometric distribution, a geometric distribution, a negative anomial distribution. In the continuous case we discussed uniform, normal distribution, exponential distribution, gamma distribution etcetera. In the bivariate case, we saw bivariate normal distribution and we saw a brief view of multivariate distributions. In the theory of several distributions then, we saw some specific distributions which are useful for carrying out inference that is, estimation and testing. So, we looked at sampling distributions such as T, chi square and F distribution.

In the next part, we have seen that when the data is given we need to represent the data in a graphical form or by tabular form. So, we saw the graphical representations and the tabular representations, the frequency classifications and certain measures of central tendency and dispersion based on this. Then, we also studied the concept of statistical inference that what we want to do in inference, so, we want to do estimation, we want to do, carry, create a confidence interval, or we want to do the testing of hypothesis. We have discussed all these methods in detail.

In the theory of estimation we considered how to derive estimators for example, using the method of moments or the maximum likelihood estimator. We also saw the criteria for judging the goodness of estimators using the criteria such as unbiasness, consistency efficiency, the criteria of minimum variance, unbiased estimators, etcetera. We gave the method of constructing confidence intervals. In the testing of hypothesis we introduced the concept of most powerful tests and how to derive them in particular, we saw the applications for deriving the test of for one sample problems and two sample problems when the data are taken from the normal populations. We also saw the test for the proportions in binomial populations.

Finally, we discussed chi square test for goodness of fit for various situation such as testing whether the data comes from a given distribution, whether we can assume independence in contingency tables, whether the proportions or responses are homogeneous when the sampling is done from different strata. Then, we also introduced two other measures of correlation or association called Spearman’s rank correlation coefficient and Kendall’s tau coefficient of association.

So, we have covered all the elementary topics of probability distribution theory estimation and testing of hypothesis. A detailed theory about testing and estimation will be covered in the course statistical inference where we tell in detail how to derive the tests, how to obtain the estimators, how to evaluate the performance of the estimators, how to find the shortest length fixed with confidence interval, or how to find highest coefficient and the fixed with confidence intervals, etcetera. Some of the other areas of the statistical inference are statistical quality control, regression analysis, multivariate analysis, designs of experiments, time series and forecasting, so, these are parts of different courses that we will be doing in upcoming courses.

So, the references as I have mentioned earlier one can, there are excellent text books on probability and statistics such a Hines and Montgomery, etcetera, there is a book by Bhattacharya and there is a book by Milton and Arnold, there is a book by Sheldonross, there is a book by uh v k Rohatgi and ((Shadily)). So, one can use any of these text books for various concepts and the problems. I have discussed in detail each concept with various applications, so, I think this would be quite helpful. With this we end this particular course of probability and statistics.