

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

Lecture - 79
Testing for Independence in rxc Contingency Table – II

(Refer Slide Time: 00:22)

Candidate	R_i Judge 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

$\sum D_i^2 = 28$

$R = 1 - \frac{6 \times 28}{20 \times 99 - 5 \times 33}$

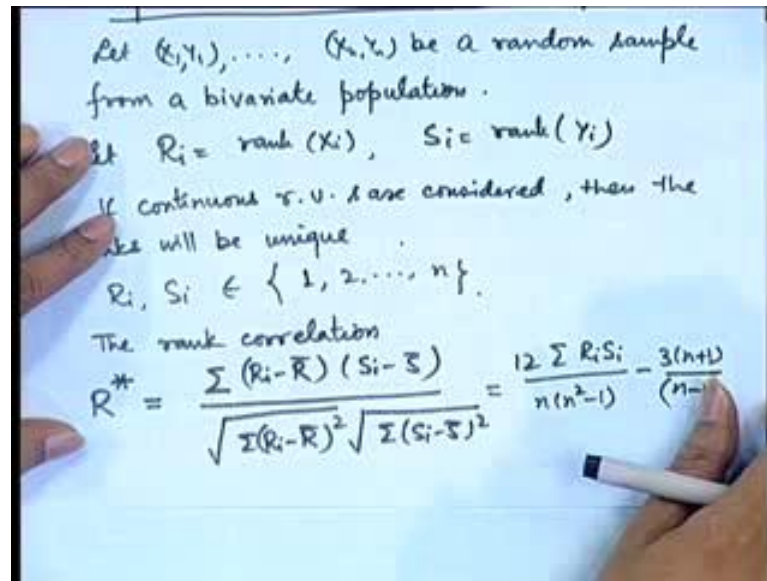
$= 1 - \frac{168 \times 28}{165}$

$= \frac{137}{165} < 1$

≈ 0.8

Now, in partial observer would like to see whether the selection has been fair or not; that means, whether the 2 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: 01:05)



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

Let $R_i = \text{rank}(X_i)$, $S_i = \text{rank}(Y_i)$

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

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

The rank correlation

$$R^* = \frac{\sum (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum (R_i - \bar{R})^2} \sqrt{\sum (S_i - \bar{S})^2}} = \frac{\sum R_i S_i - \frac{3(n+1)}{4}}{n(n^2-1) - \frac{(n+1)^2}{4}}$$

Spearman's rank correlation coefficient; so let $X_1, Y_1, X_2, Y_2, X_n, Y_n$ be a random sample from a bivariate population. Let R_i be equal to rank of X_i , and S_i be say rank of Y_i . So, what is the meaning of rank? We arrange these X_i 's and Y_i 's in a increasing order. So, by smallest value is given rank one, the second smallest is given rank 2 like that. So, now, the data may be discrete or continuous, if the data is continuous then the probability of equality of 2 observations will be 0. So, in that case the rankings will be unique; that means, all the ranks from the 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, R_i, S_i both will belong to the set 1 to n . If we want to calculate the whole relation coefficient between ranks here, I will give a notation say R^* , because R we used for Karl Pearson correlation coefficient. So, this will be $\sum (R_i - \bar{R})(S_i - \bar{S})$ divided by $\sqrt{\sum (R_i - \bar{R})^2} \sqrt{\sum (S_i - \bar{S})^2}$. So, basically what we are doing in place of X_i and Y_i 's we are using R_i 's and S_i 's. 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 this 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 R i, this will be the some of the numbers 1 to up to n; similarly S i's they will be 1 to n therefore, all these quantities can be actually evaluated.

(Refer Slide Time: 04:45)

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

$$\bar{R} = \bar{S} = \frac{n+1}{2}$$

$$\sum R_i^2 = \sum S_i^2 = \frac{n(n+1)(2n+1)}{6}$$

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

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

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

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

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

So, we can see here sigma R i or sigma S i, 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 R i minus R bar whole square, or sigma S i minus S bar whole square, because sigma R i square or sigma S i 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 ten 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 R i S i 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 R i S i, 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 d i to be the differences in the, that you write capital D i differences in the ranks R i minus S i. So, this of course, you can write as R i minus R bar, minus S i minus S bar, because R bar and S bar are same. So, they will cancel out.

So, if I look at sigma of D i square, this terms out to be sigma R i minus R bar square plus sigma S i minus S bar whole square, minus twice sigma R i minus R bar into S i

minus \bar{S} . So, this terms out to be $\frac{1}{n} \sum (R_i - \bar{R})(S_i - \bar{S})$ divided by $\sqrt{\frac{1}{n} \sum (R_i - \bar{R})^2} \sqrt{\frac{1}{n} \sum (S_i - \bar{S})^2}$. So, from here $\sum (R_i - \bar{R})(S_i - \bar{S})$ can be substituted in terms of $\sum (D_i)^2$. So, if you substitute that here in this particular formula, we get R^* is equal to $\frac{1}{n} \sum (D_i)^2$ divided by $\frac{1}{n} \sum (R_i - \bar{R})^2$ minus 1. So, this particular term is called a Spheremans 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.

So, as an example let me calculate in this particular case. So, here these can be considered as R_i , this can be considered as D_i S_i . So, let us calculate here D_i 's that is $R_i - S_i$. So, this is 0, this is 1, 1, minus 2, 1, 1, minus 3, minus 1, 3 and 1. So, if we look at D_i square that is 0 1 1 4 1 1 9 1 9 1. So, $\sum D_i$ square is equal to 1 plus 1 plus 4, 28. So, now, if we substitute in this particular formula that is $\frac{1}{n} \sum (D_i)^2$ divided by $\frac{1}{n} \sum (R_i - \bar{R})^2$ minus 1. So, the value of R^* that is equal to $\frac{1}{n} \sum (D_i)^2$ divided by $\frac{1}{n} \sum (R_i - \bar{R})^2$ minus 1, which can be simplified little bit $\frac{28}{165} - 1$. So, that is equal to $\frac{168}{165}$ some value will come 3 this is 5 this is 33.

So, we get here $\frac{168}{165}$ divided by 1 I am sorry, this is $\frac{28}{165}$ that is equal to $\frac{145}{165}$, 137 by 165. So, which is of course, less than 1, but it is near about if you 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 2 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; that means, we will have in general minus 1 less than or equal to R^* , 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 you are considering R^* 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 correlation 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 we will just consider one more example.

(Refer Slide Time: 12:31)

Example :

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

R _i	5	3	7	4	8	2	10	6	1	12	9	11
S _i	6	2	8	3	7	4	11	5	1	10	9	12
D _i	-1	1	-1	1	1	-2	-1	1	0	2	0	-1

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

$\sum D_i^2 = 16$

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

There is a high degree of relationship

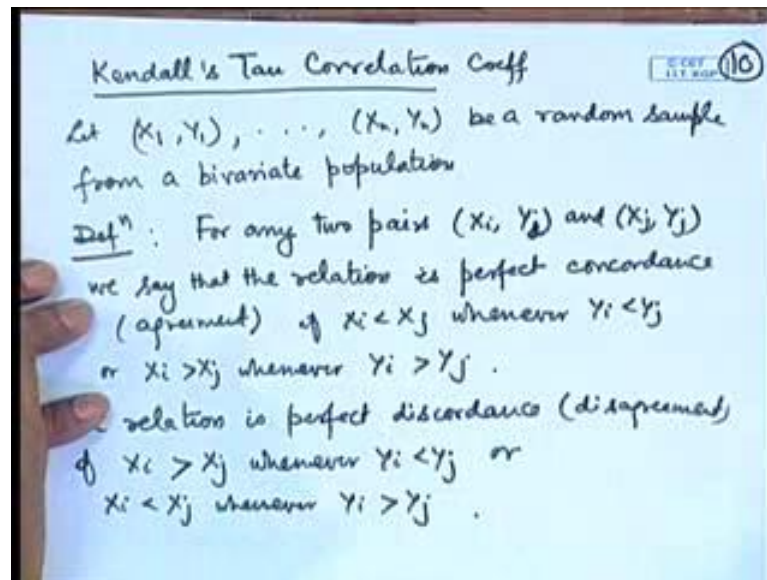
So, we have heights of fathers and heights of say sons and data is as follows sorry this is not heights this is weights. So, there are 12 pairs 65 68 63 64 67 70 64 65 68 69 62 66 70 75 66 67 61 63 72 74 69 71 71 76.

Now, rather than calculating the correlation between the raw values of the weights, we will consider their ranks. The reason is that you 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 R i's and S i's here, it terms 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, there is a highest the heaviest father has the heaviest son, 12, 10, 9, 9, 11, 12 we want to see the concordance or not. So, if you look at the R star for that we need to calculate the differences D i's 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 D i square that is equal to 16. So, 1 minus 6 into sigma D i 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. Yet another measure of correlation where in place of the raw values we look at only their magnitudes is Kendall's tau correlation coefficient.

(Refer Slide Time: 16:23)



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 site; that means, if you are having 2 value say X_i 's and X_j then if X_j is on a higher side of X_i then whether Y_j is on the higher side of Y_i or in the reverse. So, this is called concordance or accordance. So, let us write let $X_1, Y_1, X_2, Y_2, \dots, X_n, Y_n$ be a random sample from a bivariate population. So, we have the following definition for concordance or discordance for any 2 pairs X_i, Y_i and X_j, Y_j . We say that the relation is perfect concordance. So, concordance here means agreement, they are in the agreement if X_i is less than X_j , whenever Y_i is less than Y_j or X_i is greater than X_j , whenever Y_i is greater than Y_j . The relation is perfect discordance or disagreement if X_i is greater than X_j , whenever Y_i is less than Y_j or X_i is less than X_j whenever Y_i is greater than Y_j .

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

(Refer Slide Time: 19:22)

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

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

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

Defⁿ $\tau = \pi_c - \pi_d \rightarrow$ Kendall's Tau
Correlation Measure of association

If the distⁿ 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 π_c is equal to probability of $X_j - X_i$ into $Y_j - Y_i$ greater than 0. You can see this is the probability of concordance and we can consider π_d as the probability of discordance of course, we may have to include equality at one place to exhaust all the possibilities. So, if the random variables are continuous, then you will have $\pi_c + \pi_d = 1$, because the probability of equal to 0 will be negligible or 0.

So, we define tau that is equal to $\pi_c - \pi_d$ 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 $\pi_c - 1$, by putting $\pi_d = 1 - \pi_c$ or you can say it has $1 - 2\pi_d$. Now if X and Y are continuous and independent, then by symmetry you must have probability of $X_i < X_j$ is equal to probability of $X_i > X_j$ that is equal to half.

(Refer Slide Time: 21:35)

Then

$$\begin{aligned} \pi_c &= P(X_i < X_j) P(Y_i < Y_j) + P(X_i > X_j) P(Y_i > Y_j) \\ &= P(X_i > X_j) P(Y_i < Y_j) + P(X_i < X_j) P(Y_i > Y_j) \\ &= \pi_d. \end{aligned}$$

ie for independent & continuous r.v.s.

$$\tau = 0$$

The converse is not true.

in bivariate normal populations:

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

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

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

So, in that case this π_c value will be equal to probability of X_i less than X_j into probability Y_i less than Y_j plus probability of X_i greater than X_j into probability of Y_i greater than Y_j , and because of the symmetry this can be then written as a X_i greater than X_j into probability of Y_i less than Y_j , that is probability of X_i less than X_j into probability of Y_i greater than Y_j which is nothing, but π_d .

So, that is for independent and continuous random variables τ 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 τ is 0, if and only ρ is 0 we look at a sample measure for or you can say sample estimate for τ . So, we can define like this, define a ψ function for x_1, y_1 and x_2, y_2 pair as 1 if they are in concordance otherwise you define it to be 0.

Then naturally expectation of ψ is equal to π_c . So, we can consider an average of the all such pairs.

(Refer Slide Time: 24:14)

Take $U = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \psi((X_i, Y_i), (X_j, Y_j))$ (13)

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

$H_0: \tau = 0$
 $H_1: \tau \neq 0$

For large n , $\frac{3\sqrt{n(n-1)}}{\sqrt{2(2n+5)}} T \xrightarrow{L} N(0,1)$ under H_0

Approximation is good for $n \geq 8$.

That means we may consider U is equal to 1 by $n \times 2$ for all $n \times 2$ pairs of X_i, Y_i and X_j, Y_j for $1 \leq i < j \leq n$; then T that is equal to $2U - 1$ 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 τ that is $H_0: \tau = 0$, against $\tau \neq 0$.

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 $3\sqrt{n(n-1)} / \sqrt{2(2n+5)} T$ is asymptotically normal $N(0, 1)$ under H_0 . So, one can use this for testing the hypothesis $\tau = 0$ and it has been observed that approximation is good for n greater than or equal to 8. Let me give an overview of the topics that we have covered in this particular course till now, we have 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 (Refer Time: 26:42) 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, base 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 characteristic such as mean, variance third order moments, fourth order moments, a moment generating function measures of symmetry, measures of peakedness, which are called measures of a skewness and kurtosis.

We also saw some sort of probability bounds in the form of semi shapes inequality, we discussed the jointly distributed random variables; that means, in place of 1 if we have 2 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, Poisson 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 hypergeometric distribution, negative geometric distribution, a negative binomial 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 mu 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 distribution 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 these.

Then we also studied the concept of a statistical inference, that what we want to do in inference; so we want to do estimation, we want to do can 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 unbiasedness, 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 tests of hypothesis 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 or homogeneous, when the sampling is done from different strata.

Then we also introduced 2 other measures of correlation or association called Spheremans 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 width confidence interval or how to find highest coefficient and the fixed width 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 test books on probability and a statistics such as Hines and Montgomery's etcetera; there is a book by Bhattacharya and Jonson; there is a book by Milton and Arnold; there is book by Sheldon Ross; there is a book V K Rohatgi and Salley. So, one can use any of these test 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.