

Statistical Learning for Reliability Analysis
Dr. Monalisa Sarma
Subir Chowdhury of Quality and Reliability
Indian Institute of Technology – Kharagpur

Lecture – 39
Correlation Analysis (Part – II)

Hello guys. So, in continuation of our discussion on relation analysis, we will be the last class basically we have discuss Pearson correlation analysis and we will start a class from that point onwards.

(Refer Slide Time: 00:42)



So, in this class we will be discussing Spearman correlation analysis and chi square test, these are the 2 different correlation analysis where Spearman correlation analysis we have already seen in the last slide that we use for ordinal data and chi square test we use for nominal data. The Spearman correlation analysis is also called rank correlation coefficient, because why because Spearman correlation less is as I told you, it is applicable for ordinal data ordinal data means there is a ranking among the data. So, that is why it is also called rank correlation coefficient.

(Refer Slide Time: 01:16)

Charles Spearman's Correlation Coefficient

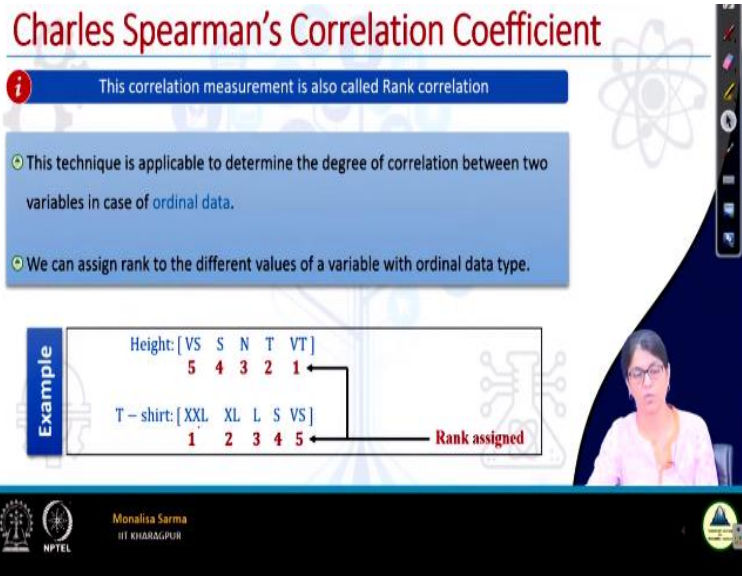
i This correlation measurement is also called Rank correlation

- This technique is applicable to determine the degree of correlation between two variables in case of ordinal data.
- We can assign rank to the different values of a variable with ordinal data type.

Example

Height:	[VS	S	N	T	VT]
	5	4	3	2	1
T-shirt:	[XXL	XL	L	S	VS]
	1	2	3	4	5

Rank assigned



So, now, if we can like see this example, the height of a person and the size of the garment so, if you are interested in finding out the relation between is any association between height of a person and the size of a garment is there any association between these 2 and if it is there, what sort of association, what is the strength of the association? So, as already mentioned, here this technique is applicable to determine the degree of correlation between 2 variables in case of ordinal data.

Now height if we can consider it as a interval I mean we can consider the ratio data also now see the ratio data we can convert a ratio data to ordinal data some likewise, I want to give you one more example here likewise the size of the house. The size of the houses is very much a numeric variable that is 1200^2 feet 1400^2 feet 2000^2 feet 3000^2 very much a numeric variable. This again, we can convert it to ordinal data how.

If we can convert the size into different categories, which is width the interesting ranking among them. So, that was similarly now height I have converted to an ordinal data like how this is so starting from very tall, tall, normal height, short, very short so, it is there is an ordering among this data. Similarly, this size of the garments are very small, small, large, extra large extra extra large.

So, now, when we design this way, these are the 2 data these are 2 variables height and size garment size, we want to find out the association between these 2, in such case we will be using Spearman correlation coefficient. So, here one thing to be noted here, see if when we have to assign so first we will have to assign the rank to these different data different

categories. So, for the height we have different categories very short, normal tall, very tall, we will assign the rank of each category.

Similarly, for T- shirt also we will assign the rank to each category. So, now, how this rank is assigned to the highest value we assigned a rank 1 like in a class who scores the higher we give it a rank 1. So, similar the highest value we assigned it rank 1 so here are very tall is rank 1 tall is 2 similarly normal is 3 and very smallest 5. Similarly, for garment size, XXL is the highest so we give it rank 1 and very small, we give it rank 5.

(Refer Slide Time: 03:50)

Charles Spearman's Correlation Coefficient

Definition : Charles Spearman's correlation coefficient

The rank correlation can be defined as

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where

d_i = Difference between ranks of i^{th} pair of the two variables
 n = Number of pairs of observations

The Spearman's coefficient is often used as a statistical methods to aid either proving or disproving a null hypothesis of no correlation.

X	Y
x	y

Now here to find out the correlation coefficient between 2 ordinal data type, this is a we have to find out basically the value of r_s that is called as Spearman's correlation coefficient r_s is a Spearman correlation coefficient. So, it has how the Spearman correlation coefficient what is that expression for this? This is the expression for this. So, what it is given we have a term d_i and we have a term here n , n we know is the size of the sample and what is d_i ? d_i is the difference between rank of the i^{th} pair of the 2 variables.

So X has different values x_1 to x_n Y has different values y_1 to y_n . So d_i is the difference between the i^{th} pair difference between x_i and y_i we have assigned rank to each variable. So difference in rank of the i^{th} pair difference in rank to the first pair means what is the difference in rank of y_1 and x_1 difference in rank of the second pair means difference in rank of x_2 and y_2 . So d_i is the difference rank between ranks of i^{th} pair of the 2 variables. So n is equals to number of pairs of observation.

And like Pearson correlation also the Spearman coefficient is often used as a statistical method to aid either proving or disproving a null hypothesis of no relation here also we will consider a null hypothesis that there is no correlation. Now, Pearson correlation analysis what we have seen that is very much a parametric analysis. The data that we have considered has to come from a normal populations. So, that is a Spearman correlation is a parametric analysis.

Now, Spearman you see, when the data is or ordinal from that only it makes it clear that it cannot be parameterised and of course, it is not parametric here we are not bothered about the populations parents, the data has come from what sort of populations it is a very robust is we are not at all bothered about the type of the population from which where we got the data there is a nonparametric analysis.

(Refer Slide Time: 05:49)

Charles Spearman's Coefficient of Correlation

Example:

The hypothesis that the depth of a river **does not progressively increase** with the width of the river.

A sample of size 10 is collected to test the hypothesis, using Spearman's correlation coefficient (shown in the table.)

Sample#	Width in m	Depth in m
1	0	0
2	50	10
3	150	28
4	200	42
5	250	59
6	300	51
7	350	73
8	400	85
9	450	104
10	500	96

Monalisa Sarma
IIT KHARAGPUR

So, now, let us discuss this Spearman correlation coefficient using an example. So, what example is given here the hypothesis that a depth of a river does not progressively increase with the width of the river. So, I am hypothesizing that depth of the river does not progressively increase the width of the river means as the width increases height does not increase there is no association between the height and width of the river, it is not mandatory that mean width increases height increases, there is no association that is my hypothesis.

But I will have to prove that is it correct that means my null hypothesis is there is no correlation between height and width of the river. My alternate hypothesis is there is correlation means as width increases height increases. So, to prove this what I have taken? I

have taken a sample of size 10 to test the hypothesis using Spearman correlation coefficient so, these are the different data that we have taken.

Width in meter width in first data we have taken 0 0 height 0 width is not it is 0 actually because 0 does not carry any meaning here, how what can be 0 return 0 that basically by 0 what I meant to say is a very negligible width and very negligible depth, maybe you can consider channel so very negligible width very negligible depth, which I am considering for simplicity is a let me consider it as 0. So these are the different width of the river, these are the different depth of the river.

(Refer Slide Time: 07:21)

Charles Spearman's Coefficient of Correlation

Step 1: Assign rank to each data.

It is customary to assign rank 1 to the largest data, and 2 to next largest and so on.

Data	20	25	25	25	30
Assign rank	5	4	3	2	1
Final rank	5	3	3	3	1

Note:
If there are two or more samples with the same value, the mean rank should be used.

Monalisa Sarma
IIT KHARAGPUR

Now I will have to rank it, how will it rank? Height we will get the now not only that, there is 1 more thing to be here. So, of course, it is customary to assign rank on to the largest data that we have already discussed 2 to the next largest and so on. Now, what happens if there are same type of data like you see here the example here, so, it starts from the 30 and we have 25, 25, 25, 3 25 and then 20. So, what the same the definitely 30 will get rank 1 then what about this 3 25 it will get rank 2 all will get rank 2 or will get 1 will get 2, 3 and 4.

So what happens is that first we assigned a rank in like this first 25 I have got 2 second 25 I have got 3 third 25 I have got 4 and then we find a mean of this. So mean of this 3 then rank 3 assigned to all this data and 20 is 5, if there are 2 or more symbols with the same value mean ranks should be used.

(Refer Slide Time: 08:15)


Charles Spearman's Coefficient of Correlation

Step 2: The contingency table will look like

Sample#	Width	Width rank	Depth	Depth rank	d	d ²
1	0	10	0	10	0	0
2	50	9	10	9	0	0
3	150	8	28	8	0	0
4	200	7	42	7	0	0
5	250	6	59	5	1	1
6	300	5	51	6	-1	1
7	350	4	73	4	0	0
8	400	3	85	3	0	0
9	450	2	104	1	1	1
10	500	1	96	2	-1	1

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \times 4}{10 \times 99} = 0.9757$$

$$\sum d^2 = 4$$



So, now for this example the height and width of the river. So, that accordingly I will rank the highest value here you see highest value is 500. So I have rank 1. Similarly, you see in the depth the highest value is 104. So 104 is rank 1. So, accordingly I have rank all the data width I have some rank depth I have given some rank, when the width is 0 depth is 0 and this total there are 10 data and this is the 10 rank. So when width 50, depth is 10 this rank is 9 and this rank is 9. So, what will be d²? d² will be x² - y², so 9 - 9 is 0.

So, similarly, I found a d_i value for all this data, like see here example here, for 250 it is the width rank is 6 and it is depth rank is 5. So, what will be the d_i? d_i will be 1, 6 - 5 = 1. Similarly, here for 300 this is 5 this is 6, so, it will be -1. So, this is my d_i data I found out the d_i data then I found this square then this is my formula which we have already seen. I put the value and I got this 0.9757. So, 0.9757 is my r_s value.

Now this 0.9757 is it really significant value that we will have to find out. Again similarly, the way we have done significant test for the Pearson similarly here we also will do the significant test for Pearson.

(Refer Slide Time: 09:44)

Charles Spearman's Coefficient of Correlation

Step 3: Spearman's rank significance table

To see, if this r_s value is significant, the **Spearman's rank significance table** (or graph) must be consulted.

Note: The degrees of freedom for the sample = $n - 2 = 8$
With the significance level 1% , null hypothesis is rejected.

Monalisa Sarma
IIT KHARAGPUR

So 0.9757 likes for Spearman we have in the any standard textbook you will find Spearman and graph in this sort of a there is a Spearman graph basically and for other T table normal table we get a table of values. Spearman we get this sort of graph it will in any standard statistical book. So, here you see if now, for this example how many data's are there? There are total 10 data, 10 data means what is the degrees of freedom $10 - 2 = 8$. So, 8 degrees of freedom.

Suppose, we are interested in finding at 1% significance level that means, 1% significance level means 0.1 you see the 1% significance level is that this can you see this is the graph this is the 1% significance level. So, if my value lies above this then it is in a critical region if my values lies below this graph then it is an acceptance region. Now, what is my degrees of freedom? Degrees of freedom is 8. So, 8 this is my 8 and what is my value? My value is 0.9757 so, for 8 0.9757 will be somewhere here it is very much above the graph.


So, it is very much above the graph for 1% that means, it falls in a critical region it falls in a rejection region. If it falls in the rejection region that means, what my null hypothesis is rejected that means, there is what is say correlation there is an association between height and depth of the river.

(Refer Slide Time: 11:08)


Charles Spearman's Coefficient of Correlation

Step 4: Final conclusion

- From the graph, we see that $r^2 = 0.9757$ lies above the line at 1% significance level with degrees of freedom 8. ✓
- Hence, there is a **greater than 99%** chance that the relationship is significant (i.e., not random) and hence **the null hypothesis should be rejected.** ✓
- Thus, we can reject the null hypothesis and conclude that in this case, **depth of a river progressively increases the further with the width of the river.**



Monalisa Sarma
IIT KHARAGPUR



So, far we see that lies above the line at 1% significance level with the degrees of freedom 8 hence, there is a greater than 99% chance that a relationship is significant that means it is not just randomly we got this data and the hence the null hypothesis should be rejected. Thus we can reject the null hypothesis and conclude that in this case depth of a river progressively increases the further with the width of the river.

So, before starting with the chi square correlation analysis. So, I want to mention one more thing here. As I told you our Pearson correlation analysis is very much a parametric method Spearman correlation analysis is very much a nonparametric method. So, when we use Pearson? Pearson we use when the data is linear. Spearman we use when the data is monotonic we have seen in the last class remember this.

When it is linear if we can graph a line monotonic we have seen it may increase in a relatively maybe increase in the same direction or in the opposite direction, but may not be in a constant rate as what we can see in linear. So, basically we may get some sort of slight curve we may get it may increase 1 increases other increases, but it may not be at a constant rate then we call it a monotonic relation, so Spearman correlation analysis we use when the relationship is monotonic.

Now, the question is how do we know that whether it is linear or monotonic? Suppose, we have plotted from given the data we have plotted the scatter plot it. Now, from the scatter plot if we can very well understand what does this scatter plot indicates it does it indicates a linear

or at negative monotonic if we are not very sure whether it is monotonic or linear, we are not very sure of it, then it is better we go for the Spearman correlation analysis.

Definitely, if it is linear, then we use Pearson will give us more results, because when we are in a numeric data when we are using Spearman that means what we have to be this numeric data we have to convert it to ordinal data, when we are converting to ordinal data, we are losing some information. So, that way so, in a linear relationship, if using Spearman, then of course, we lose some information that way our results will not be very accurate, but still we will get a result which is closer to Pearson only.

But if it is not linear and we try using linear we will lose lots of information we will not get the correct information at all. So, from the graph if we see that we are not very sure it is linear or monotonic, better to go for Spearman, result may not be as precise as Spearman, but we are not taking the risk because if it is monotonic and we are using Pearson we will not get a good result. That means what if it is Spearman can be used for linear relationship also.

It is mainly used for monotonic relationship it can be used for linear relationship also, that means, when the data it very much violates the linear relationship we go for Spearman relationship, but it can also be used for linear relationship with precision becomes lesser. So now next we will discuss is the chi square correlation analysis. This is also a very much nonparametric analysis very robust way we are not considered about the distribution of the parent populations.

We do not have to bother about the variance of the populations how it is whether the variance of the different population is same or not nothing we not bothered anything. So, as I told you chi square correlation analysis we use for nominal data where we have different categories of data.

(Refer Slide Time: 14:49)

Chi-Squared Test of Correlation


i This method is also alternatively termed as Pearson's χ^2 -test or simply χ^2 -test

- This method is applicable to categorical data only.
 - Suppose, two attributes A and B with categorical values

$$A = a_1, a_2, a_3, \dots, a_m \text{ and}$$

$$B = b_1, b_2, b_3, \dots, b_n$$
 having m and n distinct values.

A	a_1	a_2	a_3	a_4	a_5	a_6	...
B	b_1	b_2	b_3	b_4	b_5	b_6	...
 - Between A and B, we are to find the correlation relationship, if any.


 Monalisa Sarma
 IIT KHARAGPUR

Like take an example here suppose there are 2 attributes A and B with categorical values. Let me give an example what sort of example? Suppose in a population there are different categories, we can divide the whole people in a particular population, you can divide into 3 categories one, the person who had severe COVID, one group of persons who had mild COVID and another group would not have COVID. So help if I consider this in my best variable as health status of the people of the population.

So I will categorize it into 3 different parts. So, let me say this is a 1, a 2, a 3, what is a 1? a 1 is the people who had severe COVID, a 2 is the people who had mild COVID, a 3 is the people did not have COVID. So A is the health status of the people and the populations. So this attribute I have 3 different categories. Similarly, b may be again, I am classifying this population B into 2 different categories based on the vaccination status.

So my attribute is the name maybe vaccination status, so 1 is fully vaccinated. Another maybe partially vaccinated another maybe not vaccinated at all or maybe 2 categories vaccinated, not vaccinated. So, the b 1, b 2 vaccinated, not vaccinated if A is 3 categories is by fully vaccinated, partially vaccinated, not vaccinated b 1, b 2, b 3. So, this way if we have in a population if we have 2 or more attributes, which have different categorical values, then we can use chi square test of correlation.

So, how we do that? So, basically here we need to find out the correlations between A and B. So, correlations between A and B means in the example, whatever given again, my correlation is that vaccination did not have any effect on the COVID health status of the

people. So there is no correlation vaccination did not have any health status other people this is my null hypothesis and alternative hypothesis vaccination have effect on the COVID status. So that means between A and B, I tried to find a correlation is there any relationship among that?

(Refer Slide Time: 17:07)

Contingency Table

Given a data set, it is customary to draw a contingency table, whose structure is given below.

	b_1	b_2	b_j	b_n	Row Total
a_1	✓	✓					✓
a_2	✓	✓					✓
⋮							
a_i	✓	✓					✓
⋮							
a_m	✓	✓					✓
Column Total	✓	✓					✓
							Grand Total

So, how do I do that first and for chi square test? I draw a contingency table. What is the contingency table? And here in the rows, I have all the A values, all the categories, all the groups of the A attribute. So what was my attribute? My A attribute maybe had severe COVID, mild COVID, had no COVID. So, these are my different groups. So this I have the different rows for that. And similarly, my columns are the vaccination status, not vaccinated, vaccinated. So b_1, b_2 , maybe here we have m b_1 to b_n similarly, a_1 to a_m .

So, I will draw the contingency table this way that is my first step. Second step is then I will find a row total and I will find a column total. What does row total indicates? Row total indicates all the sum of all this. What does this indicate? This indicates this is suppose b_1 is vaccinated, vaccinated people how many vaccinated people got severe vaccine COVID, minor COVID and no COVID this total will give me my column total.

Now, this column total non vaccinated people had severe COVID, mild COVID, no COVID so, this will give me my column total this is also called residual column this is column residual. Similarly, I can find out row total what is row total? People who have severe COVID how many of them are vaccinated, how many of them are non vaccinated, this total as I got is row total. Row total is also called row residual.

Similarly, here I will find what to say vaccinated people having mild COVID non vaccinated people having mild COVID some of this I will get it here that is row total, similarly and get a row total here. So that is how I first draw my contingency table? I have drawn the contingency table. Now, I have to draw the contingency before filling out the row table, I will have to fill these values, I will have to fill as well as within this column, in this continuance is it.

(Refer Slide Time: 19:13)

χ^2 - Test Methodology

Entry into Contingency Table: Observed Frequency

⊙ In contingency table, an entry O_{ij} denotes the event that attribute A takes on value a_i and attribute B takes on value b_j (i.e., $A = a_i, B = b_j$).

A	a_1	a_2	a_3	a_i	a_5	a_i	...
B	b_1	b_2	b_3	b_j	b_5	b_j	...

	b_1	b_2	b_j	b_n	Row Total
a_1	O_{11}						
a_2							
⋮							
a_i				O_{ij}			
⋮							
a_m							
Column Total							Grand Total

How to fill this table before going to the row total first so to fill this total filling this total is any entry O_{ij} what does this indicate? O_{ij} indicate this $A = A_i, B = B_j$. So, in my example, what does a_1, b_1 indicates a this is O_{11} this would be what this O_{11} indicates here O_{11}, O_{11} indicates people who are vaccinated and got severe COVID so, what is this? This is O_{21} . What is O_{21} indicates? O_{21} indicates people who were vaccinated and get got mild COVID.

This is people who are vaccinated and got severe COVID. So, this way I fill this data. So, this is how I fill the whole table and then I find out the row total and column total. So, in their contingency table and entry O_{ij} denotes the event that attributes A takes on value A_i and attribute B takes in a value of B_j .

(Refer Slide Time: 20:29)



χ² - Test Methodology


Entry into Contingency Table: Expected Frequency

⊙ In contingency table, an entry e_{ij} denotes the expected frequency, which can be calculated as


$$e_{ij} = \frac{\text{Count}(A=a_i) \times \text{Count}(B=b_j)}{\text{Grand Total}} = \frac{A_i \times B_j}{N}$$

	b ₁	b ₂	b _j	b _n	Row Total
a ₁	0						✓
a ₂							
⋮							
a _i				e_{ij}			A _i ✓
⋮							
a _m							
Column Total	0			B _j ✓			N



Monalisa Sarma
IIT KHARAGPUR



So, now, next is after finding the O_{ij} and after filling up the whole contents table than finding out the row residual and a column residual next step is that I will find the expected frequency. How do I find the expected frequency? Expected frequency is this basically the row residual and column residuals divided by the total number of values that is the expected frequency. So, this is $A_i \times B_j / N$ for each value. If I am interested in finding this, so, it will be this value what is $a_1 \times b_1 / N$ that is how I will find out the e_{ij} ? e_{ij} is called the expected frequency.

(Refer Slide Time: 21:15)


χ² - Test


Definition : χ²-Value

The χ² value (also known as the Pearson's χ² test) can be computed as follows:


$$\chi^2 = \sum_{i=1}^m \sum_{j=1}^n \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

where o_{ij} is the observed frequency
 e_{ij} is the expected frequency





Monalisa Sarma
IIT KHARAGPUR



So, once I have the O_{ij} once I can have the e_{ij} then I can find out the chi square value. So, how is the chi square value? This is the formula for finding out the chi square value. So, for each table how do I get the chi square value? I have the O_{ij} value for each column cell I have the e_{ij} value for each cell. So, how do I find out a chi square value? O_{ij} value for each O_{ij} - this formula $(O_{ij} - e_{ij} / e_{ij})^2$ that is the chi square value for 1 cell. So, when I am interested

in finding out the chi square value of whole cell so, some of all cell value will give me my chi square.

(Refer Slide Time: 21:56)

χ^2 -Test

- The cell that contribute the most to the χ^2 value are those whose actual count is very different from the expected.
- The χ^2 statistics tests the null hypothesis that A and B are independent.
- The test is based on a given significance level and with (n-1) × (m-1) degrees of freedom, for a contingency table of size n × m.
- If the hypothesis can be rejected, then we say that A and B are statistically related or associated.

Monalisa Sarma
IIT Kharagpur

So, then again we will have to do find us what to say significance test whether the whatever value I got whatever value for chi square test I got, whether it is really significant whether the null hypothesis is accepted or null hypothesis is rejected. So, similarly, I will like in the previous method I will be doing a significance test what is the significance test? First in the cell that contributes the most to the chi square value or those whose actual count is very different from the expected when I get the chi square value for each cell.

If you take an example and if you calculate the chi square value for each cell will get different chi square value maybe same may be different for each other cell in the contingency table the cell which has the highest chi square value it is because it is actual count is very different from expected what is O_{ij} ? O_{ij} is the actual count actual even so, and e_{ij} is the expected value when the actual count is very much different from the expected value then we get a high chi square value.

So, chi square statistics test the null hypothesis that A and B are independent, this test is based on a given significant level with $n - 1 \times m - 1$ degrees of freedom for a contingency table of size n cross m. If the hypothesis can be rejected, then we say that A and B are statistically related or associated. If we can reject the hypothesis, if it falls in the rejection region, then we can say A and B are statistically related they are not independent.

(Refer Slide Time: 23:33)


χ^2 - Test


Example : Survey on Gender versus Hobby.

Suppose, a survey was conducted among a population of size 1500. In this survey, gender of each person and their hobby as either "book" or "computer" was noted. The survey result obtained in a table like the following.


GENDER	HOBBY
.....
.....
M	Book
F	Computer
.....

We have to find if there is any association between Gender and Hobby of a people, that is, we are to test whether "gender" and "hobby" are correlated.





Monalisa Sarma
IIT KHARAGPUR



So, we will see with the help of an example a very simple example will also do some example in tutorials. So, do not worry, this is a bit critical. So, suppose a survey was conducted among a population of size 1500 total population sizes 1500 and this survey is gender of each person and their hobby as either books or computer was noted. So, what is the survey? Survey the gender of each person is noted and a hobby is either a book or computer we are making the problem very simple.

There we are just considering 2 gender there is male and female and we are considering just 2 hobbies book and computer there may be many other hobbies but we as a sponsor as if people hobbies and reading books and walking on computer these are the 2 hobbies to make the problem simpler we have just considered that. We have to find if there is an association between gender and hobby of a people that is we are to test whether gender and hobby are correlated.

(Refer Slide Time: 24:29)


χ²-Test


Example : Survey on Gender versus Hobby.

From the survey table, the **observed frequency** are counted and entered into the contingency table, which is shown below.

GENDER	HOBBY
—	—
—	—
M	Book
F	Computer
—	—

		GENDER		
		Male	Female	Total
HOBBY	Book	250	200	450 ✓
	Computer	50	1000	1050 ✓
	Total	300	1200	1500




 Monalisa Sarma
 IIT KHARAGPUR

So from the survey table, the observed frequency are counted and entered into the contingency table. From the table we will have to now first develop that contingency table. So this is the contingency table we have developed. So total 250 male has given reading books as a hobby and 200 female has given reading books as a hobby. Similarly, 50 male has given working on computer the hobby 1000 female has given working on computer the hobby so, this is the observed frequency.

This is what we have observed from the survey this is called observed frequency that means filling up the O_{ij} values now that they O_{ij} values. So, now we will find a row total and a column total that is row residual and a column residual, what is the row residual for this is addition of this 450. Similarly, this is 1050 and column residuals, adding this and adding this. Now, next what we will find? We will find the expected frequency. How do I find the expected frequency?

Expected frequency for this is $450 \times 300 / 1500$ for this is $450 \times 1200 / 1500$ for this value, for this value is $1050 \times 300 / 1500$ for this value is $1050 \times 1200 / 1500$. That is how I will find out the expected frequency.

(Refer Slide Time: 25:53)


χ² - Test

Example : Survey on Gender versus Hobby.

From the survey table, the **expected frequency** are counted and entered into the contingency table, which is shown below.

GENDER	HOBBY
---	---
M	Book
F	Computer
---	---

HOBBY	GENDER		
	Male	Female	Total
Book	90	360	450
Computer	210	840	1050
Total	300	1200	1500



So, this is the expected frequency value I got from the observed frequency.

(Refer Slide Time: 26:00)

χ² - Test


Using equation for χ² computation, we get

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$

$$= 507.93$$

	Male	Female
Book	250 (90)	200 (360)
Computer	50 (210)	1000 (840)

- This value needs to be compared with the tabulated value of χ² (available in any standard book on statistics) with degrees of freedom (for a table of m × n, the degrees of freedom is (m - 1) × (n - 1); here m = 2, n = 2).
- For 1 degree of freedom, the χ² value needed to reject the hypothesis at the 0.01 significance level (i.e., 1%) is 6.635.
- Since our computed value is above this, we reject the hypothesis that "Gender" and "Hobby" are independent and hence, conclude that the two attributes are **strongly correlated** for the given group of people.



So, here in this table I am showing this is the observed frequency this is the expected frequency. Now, I have the observed frequency at the expected frequency then I can find out the chi square value, chi square value of each component of each cell, if I am interested in finding out the chi square value of this cell what will be its (250 - 90)² / 90. So, this is the chi square value for this. This is the chi square value for this, this 1.

This is the chi square value this 1 is the chi square value for this 1 this cell this value is the chi square value for this cell adding all together I got the chi square value. So, I got a chi square value of 507.93. Now, what is my degrees of freedom, what is my m and n? Both m and n is 2. So, n - 1 x m - 1 so, it is 1 + 1 is my degrees of freedom is 1. So, I will see their consult the chi square table where degrees of freedom 1 and significance levels 1%.

If I considered a chi square table for degrees of freedom 1 and significance level of 1% I will get the value 6.635 you can see from the chi square table. Now, what I value I got 507.93 is very bigger. Of course, this is a very toy example that is why I have got this different result. So, 507.93 is very much bigger than this. That means it falls in the rejection region. That means hobby and gender is there is a correlation between hobby and gender.

Since our computer value is above this, we reject the hypothesis that gender and hobby are independent and hence conclude that the 2 attributes are strongly correlated for a given group of people so it is strongly correlated.

(Refer Slide Time: 27:44)

Significance Test for χ^2 -Test

Cramer's V Test

- For χ^2 -test, the most commonly used test to measure the strength of the relation is Cramer's V test. The test takes the following form:

$$V = \sqrt{\frac{\chi^2/n}{(k-1)}}$$

- Here, n is the number of total observation, and k is the number of rows or columns, whichever is less.
- For the example case, n = 1500 and k = 2. Hence, V = 0.58.
- Thus, it is neither weak nor a strong correlation; this implies that **Gender and Hobby** are related with the degree of correlation 0.58

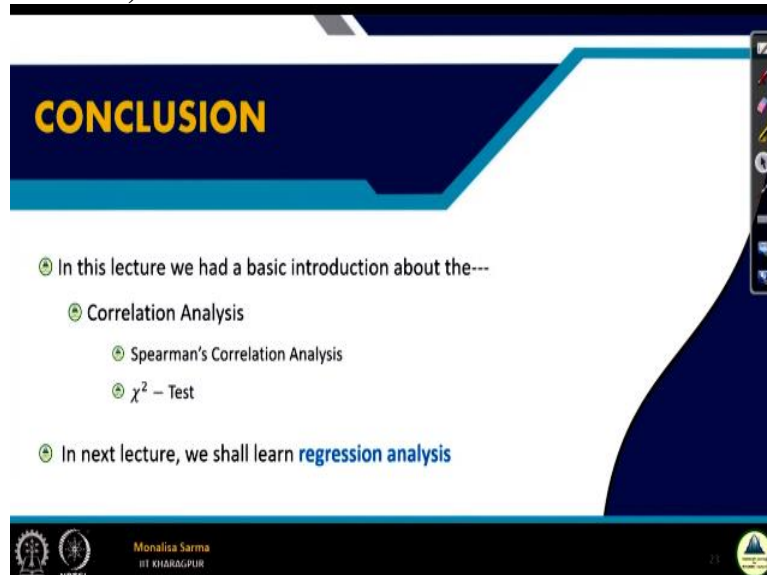
Monalisa Sarma
IIT KHARAGPUR

Now how to find out the strength of the correlation? We found it is correlated. Now how to find the strength of the correlation? So to find strength a correlation for chi square test 1 very famous test is Cramer's V test, we call it Cramer's V test. So, for chi square is the most commonly used test to measure the strength of the relation is Cramer's V test. So, the test takes this form I have the chi square value I have n and what is this k? k we can consider it a number of rows or number of column whichever is less.

What is the number of rows and number column different groups of the attributes? So, in this case, our number of rows is also 2 number of columns is also 2 and example what I have given regarding COVID my number of rows is 3 number column is 2. So then I will consider k = 2. So putting the value in this expression, I got the V value. So what is V? My V is 0.58 it is neither weak nor a strong correlation.

This implies the gender and hobby related with a degree of correlation 0.58 it is not a very strong correlation, not a very weak correlation or something very near to 0, I will call it weak correlation. So, it is not a very weak correlation also. So, we hobby and gender is correlated with a degree of correlation 0.58.

(Refer Slide Time: 29:14)



CONCLUSION

- ⊙ In this lecture we had a basic introduction about the---
 - ⊙ Correlation Analysis
 - ⊙ Spearman's Correlation Analysis
 - ⊙ χ^2 – Test
- ⊙ In next lecture, we shall learn **regression analysis**

Monalisa Sarma
IIT KHARAGPUR

So, we have discussed as Spearman correlation analysis, we are discussed chi square test and that is all for correlation analysis. In the next lecture, we will discuss regression analysis. Remember I have already told in relational analysis has either correlation analysis or regression analysis. So, in my next lecture, I will be discussing regression analysis.

(Refer Slide Time: 29:36)



REFERENCES

- ⊙ The detail material related to this lecture can be found in

The Elements of Statistical Learning, Data Mining, Inference, and Prediction (2nd Edn.), Trevor Hastie, Robert Tibshirani, Jerome Friedman, Springer, 2014.



Monalisa Sarma
IIT KHARAGPUR

So, this is the reference and thank you.