

Introduction to Probability and Statistics
Prof. G. Srinivasan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture – 09
Association between categorical variables (continued)


In this lecture we continue the discussion on Association between categorical variables.

(Refer Slide Time: 00:24)

	Airline XX	Airline YY	Total
On time	86 72%	81 81%	167 76%
Delay	34 28%	19 19%	53 24%
Total	120	100	220

YY has a better on time departure performance

Consider type of flight
Point to point hopping



We were looking at this example in the previous lecture. So, we look at one set of variables which is on time departure and delay in departure and we have 2 airlines let us say we call them XX and YY. So, this is the contingency table that explains this. So, data for 220 flights and we say that airline XX in 86 instances departed on time and 34 instances there was a delay by a few minutes and so on.

So, now let us find out an association can we say for example, whether XX or YY has a better on time departure performance. Nearly by going through this data we realize that airline XX 72 percent of the times has departed on time, while airline YY 81 percent of the times has departed on time and therefore, we might say at the moment that YY has a better on time departure performance than airline XX. So, we tried to answer one question does airline YY have a better performance.

We also want to answer another question is there an association between the airline and the performance or is there no association. So, we will define a couple of metrics for the association between the categorical variables in this lecture, but let us continue discussion on that question does YY have a better performance based on this data, yes because 81 percent of the times it seems to depart on time whereas, XX does it only for 72 percent of the times.


Now we can bring in a third variable which can help us understand or which can bring in a different perspective to the whole analysis. Now let us also bring a third variable which is let us consider the type of flight and we might call that flight into 2 types which could be a point to point flight or a hopping flight. So, in a hopping flight we mean that the flight starts at airport A goes to B and then goes to C whereas, in a point to point it goes directly from A to C and so on.

(Refer Slide Time: 02:55)

	Airline XX		Airline YY		Total
	Hopping	Point to point	Hopping	Point to point	
On time	50 81%	36 62%	45 78%	36 86%	167 76%
Delay	12 19%	22 38%	13 22%	6 14%	53 24%
Total	62	58	58	42	220

If it is a hopping flight, XX has a better performance

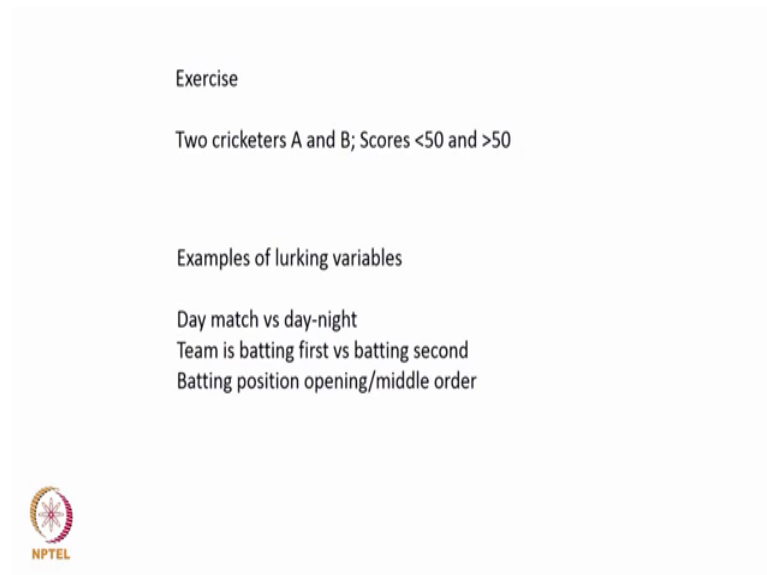
The external variable that influences the performance is called a **lurking variable** and the change is called "Simpson's paradox"



So, let us add this variable and look at this direction and if we do that and let us assume that the same data on 220 flights has now been categorized or has now been computed again based on the type of the flight. And let us say now, that the 86 flights which you can see here the 86 flights has now become 50 flights which are hopping and 36 flights which are point to point and so on. And when we do this analysis we realize that for hopping flights airline XX seem to have a better performance than YY whereas, in point to point flights YY seems to have a better performance than XX.

So, what we understand through this example is when we considered these 2 types of categorical variables we concluded that airline YY seems to have a better performance, but when we brought one more variable which is the nature or type of flight we now realize that for one type of flight XX seems to be better and for the other YY seems to be better. So, if it is a hopping flight XX seems to have a better performance so, the external variable that influences the performance is called a lurking variable and the presence or the change through this lurking variable is also called the “Simpson’s paradox”.

(Refer Slide Time: 04:34)




Exercise

Two cricketers A and B; Scores <50 and >50

Examples of lurking variables


Day match vs day-night
Team is batting first vs batting second
Batting position opening/middle order



Now, let us look at some an exercise where they can identify lurking variables and we can see the effect of them for example, we could look at the scores of two cricketers and if you want to make that comparison about consistency then we could think of what types of matches day match versus day night match, team batting first versus team batting second, batting position of the player particularly opening position, middle order position and so on.

(Refer Slide Time: 05:05)


	AA	BB	Total
≥ 30	23 46%	18 37%	41 41%
≤ 30	27 54%	31 63%	58 59%
Total	50	49	99



Now, for example, we could look at something like this greater than or equal to 30 scores two players AA and BB and we could have a performance like this.

(Refer Slide Time: 05:15)

	AA		BB		Total
	First	Second	First	Second	
≥ 30	11 48%	12 44%	14 52%	4 18%	41 41%
≤ 30	12 52%	15 56%	13 48%	18 72%	58 59%
Total	23	27	27	22	99



Whereas, if we looked at first innings and second innings then you see that the performance changes and person BB seems to now have a higher percentage than cricketer AA and so on.


(Refer Slide Time: 05:29)

Measuring association among categorical variables

Attitude towards attending classes when instructor does not take attendance

	Attend	Skip	
Fresh graduates	12	17	29
Work experience	28	15	43
Total	40	32	72

	Attend	Skip	
Fresh graduates	$\frac{29 \times 40}{72}$	$\frac{29 \times 32}{72}$	29
Work experience	$\frac{40 \times 43}{72}$	$\frac{32 \times 43}{72}$	43
Total	40	32	72



So, how do we measure association among categorical variables? So, let us take an example and try to define a measure or a metric to find out the association is there an association or is there no association. So, the data that we look at is let us look at attending classes when the instructor does not take attendance.

Now we have two types of students in the class we could have fresh graduates who had come into a master's program or and we could have people with work experience who come to a master's program. So, one categorical variable is fresh graduates and students with work experience which is shown here and the other variable is they attend classes and they skip classes.

So, let us assume we have data for 72 classes where we have we have this. So, we realize that the fresh graduates attend 12 numbers, fresh graduates who skip 17 numbers, work experience people who attend 28 numbers and work experience who skip is 43 numbers, making up for 72 number of students in the class out of which 29 are fresh graduates and 43 are people with work experience and out of these 72 let us say 40 attend classes and 32 skip classes when the instructor does not take attendance.

So, this number 72 represents the total number of people in the class and that is made up of 29 plus 43 with respect to fresh graduates and people with work experience. Now we want to know is there an association for example, is there an association between

attending the class and skipping the class we saw we fresh graduates and people with work experience.

Now, let us try to do these proportions one more time, now what we try to do is we now from this data just for the sake of computation we leave out these 4 numbers and then let us say we have total of 29 fresh graduates and 43 students with work experience with 72 students and let us say we also know this number that 40 attend classes and 32 skip classes. Now what we try to do is, to create these proportions again now the number that we have here is now treated like this, now if we if 40 out of 72 students attend now, if I take only the fresh graduates let us assume the same proportion attend and therefore, that proportion becomes $\frac{40}{72}$ into 29 and this number becomes $\frac{40}{72}$ into 43.

Now, here this number becomes $\frac{29}{72}$ into 32 and the other number becomes $\frac{43}{72}$ into 32. So, let me repeat this one more time now let us we started with this table and then let us say we found out this in the class and then we say that there are 29 fresh graduates and 43 people with work experience in a postgraduate class and we know that 40 attend and 32 skip making a total of 72 row wise total as well as column wise total.



Now we want to create these ratios again. Now, we want to create these four numbers again as ratios and let me explain how they are created. Now we say if 40 students out of 72 attend in the class proportionately how many of the 29 fresh graduates attend the class. So, that number is $\frac{40}{72}$ into 29 which is written here. Similarly if 40 out of 72 attend the class how many out of the 43 with work experience attend the class and that is given by $\frac{40}{72}$ into 43. If 32 out of 72 do not attend the class or skip the class, then how much out of these 29 skipped the class and that is given by $\frac{29}{72}$ into 32, if 32 out of 72 skip the class how much out of 43 skip the class which is given by $\frac{32}{72}$ into 43.

So, we can calculate these numbers and these numbers are calculated and shown here. So, the original data is 12 17 28 and 15 and the new calculation for example, $\frac{29}{72}$ into 40 by 72 and $\frac{29}{72}$ into 40 by 72 will be $\frac{29}{72}$ into 5, 5 8s are 40, 9 8s are 72, $\frac{29}{72}$ into 5 by 8 and that would become 16 this is rounded off.

(Refer Slide Time: 10:46)

12	17	16	13	-4	4
28	15	24	19	4	-4

Data Artificial (based on proportions) Difference

$$\chi^2 = \frac{(12-16)^2}{16} + \frac{(17-13)^2}{13} + \frac{(28-24)^2}{24} + \frac{(15-19)^2}{19}$$
$$\chi^2 = 1 + 1.23 + 0.66 + 0.84 = 3.74$$
$$Cramer's V = \sqrt{\frac{\chi^2}{n \times \min(r-1, c-1)}} = \sqrt{\frac{3.74}{72 \times 1}} = 0.23$$


So, rounded off to 16 therefore, the other one is rounded off to 13. So, that the total is 29 and once we calculate this 16 we know that this total is 40 therefore, this becomes 24 and the other becomes 19. So, we have data and then we have these proportions which are calculated and then we find the difference and in this case the difference happens to be minus 4 4 4 and minus 4.

We now calculate a number which is called chi square this is chi square ch I square. So, called chi square which is given by 12 minus 16 the whole square by 16, we had this 12 minus 16 the whole square divided by the 16, 17 minus 13 the whole square divided by this 13, 28 minus 20 4 the whole square divided by 24, which is here and 15 minus 19 the whole square divided by 19, and when we do this computation chi square becomes 3.74.

So, we calculate this number called chi square which is 3.74 M we can also calculate another number called Cramer's V and this Cramer's V is given by root over chi square divided by n into minimum of r minus 1, c minus 1. So, chi square is 3.74 divided by n is 72 r minus 1 and c minus 1 both r and c number of rows and number of columns are 2 therefore, r minus 1, c minus 1 is 2 and the minimum is 1. So, Cramer's V is root over 3.74 divided by 72 into 1 which is 0.23. So, we have now calculated 2 measures of association. So, one is the chi square and the other is the Cramer's V.

(Refer Slide Time: 13:01)

2	27
38	5

16	13
24	19


-14	14
14	-14

Data Artificial (based on proportions) Difference

$\chi^2 = 45.81$

$Cramer's V = \sqrt{\frac{\chi^2}{n \times \min(r-1, c-1)}} = \sqrt{\frac{45.81}{72 \times 1}} = 0.7976$

$V \leq 0.25$ shows weak association and ≥ 0.75 shows strong association



Now, if we look at another type of data, let us say we look at another type of data where the same 40 and 32 instead of 12 17 28 and 15 let us say we had 2 27 38 and 5. The artificial or computed proportions would still be the same as 16 13 24 and 19, let me show how this 16 was obtained. So, this 16 is actually 29 out of the 72 are fresh graduates and therefore, proportionately out of the 40 people how many are expected to attend.

So, that will be 29 by 72 into 40 which will be 16 and the other number will be 43 by 72 into 40 which is 24. So, this table will remain the same for example, those who skip will be 29 by 72 into 32 and 43 by 72 into 32 which are 13 and 19 and we would get a different kind of a difference in this and chi square would become 45.81 and Cramer's V in this case is 0.7976.

Now, the general guideline is, V less than equal to 0.25 shows a weak association and Cramer's V greater than or equal to 0.75 shows a stronger association. So, when we started with this example of 12 17 28 and 15 we calculated the Cramer's V to be 0.23. And therefore, we would say that there is a weak association or one could conclude based on this example that one cannot say with certainty that there is a strong association between the type of student and the tendency to attend or to skip.

So, V greater than equal to 0.75 shows a strong association and we can now say there is an association and one can generally conclude that people with work experience attend

classes tend to attend more classes while fresh graduates tend to skip classes. So, there is an association between the type of student and the decision to attend or skip.

(Refer Slide Time: 15:37)




Therefore we end the discussion on association between categorical variables using 2 metrics primarily the Cramer's V which comes out of the computation of chi square. So, Cramer's V less than or equal to 0.25 shows a weaker association and the Cramer's V greater than equal to 0.75 shows a strong association. For example, when we looked at the other data we had we had when we looked at this data the Cramer's V was 0.23.

If for some reason we had this kind of a data with 2 27 38 and 5 and the proportions are 16 13 24 19 I am not saying that these were obtained based on the on the earlier calculation and if we had a situation where the differences were large in this which resulted in a large value of chi square and a large value of Cramer's V then we would say in this case there is a strong association, which means one could say that people with work experience attend large number of classes while student fresh graduates would skip a large number of classes. So, now let us have a discussion on association between categorical variables.

(Refer Slide Time: 17:02)

Match the following

No.	Column A	Column B	
1	Table of cross classified counts	No association	Contingency table
2	Shown as bar chart	Cramers V	Marginal distribution
3	Measure of association between categorical variables	Contingency table	Chi squared
4	Measure of association between categorical variables (lies between 0 and 1)	Chi squared	Cramer's V
5	Produced by a lurking variable	associated	Simpson's paradox
6	Conditional distribution matches marginal distribution	cell	No association
7	Percentage within row differs from marginal percentages	Marginal distribution	association
8	Cases that match two categorical variables	Simpson's paradox	cell



And let us start with match the following. So, column a has about 8 pieces of information which have to be matched with column b. So, table of cross classified counts is called a contingency table and we have seen that. So, there is a cross classification across 2 types of categorical variables. So, shown as a bar chart so, marginal distribution can always be shown as a bar chart measure of association between categorical variables. So, in this we could either give chi square or we could give Cramer's way and the example is chi square, measure of association that lies between 0 and 1 and we can now show that the Cramer's V lies between 0 and 1 and it is a measure of association between categorical variables.


So, we get Cramer's V in this produced by a lurking variable we just now saw that Simpson's paradox is the example and that happens because of a lurking variable. When conditional distribution matches the marginal distribution then there is no association in the data, which means the proportions actually match the percentage within rows different differs from marginal percentage then there is association and cases that match two categorical variables is a cell in the table. So, this kind of helps us understand the basic concepts it also tries to tell us, what is a contingency table, what is a cell, what is a marginal distribution, what are the methods of association, and when do we have association and when we do not have association.

(Refer Slide Time: 18:48)

True or false

1. We can fill cells of contingency table from marginal counts if the variables are not associated
2. The value of chi square depends on the number of observations in the contingency table
3. Cramer's V is zero when the variables are not associated
4. The value of chi squared depends on which two variables define the rows and which two define the columns
5. If male and female are values of a variable and if the percentage female is higher, there is association between variables

True, True, True, False, False



So, let us continue with some true or false, we can fill cells of contingency table from the marginal counts if the variables are not associated true because the proportions will match. The value of chi square depends on the number of observations in the contingency table true larger the number of observations more perhaps the value of chi square will become. Cramer's V is 0 when variables are not associated when the variables are not associated the proportions are the same therefore, the difference will be 0 and therefore, chi square value will be 0 and Cramer's V will also be 0.


The value of chi square depends on which variables define the rows and which define the columns it is false, if we interchange the rows and columns the chi square value will remain the same. If male and female are values of a variable and the percentage of females is higher then there is association between variables we cannot say that it could be false because the other variable could be such that there is no association between male and female with respect to the other variable.

(Refer Slide Time: 19:59)

Question 1

- Customers were asked to give preferences for colour and shape of a product. Two teams were created by the company – each to determine the colour and shape of the product. Is it necessary to check if the two variables are associated?

Necessary to check. Good if there is no association. Otherwise the association has to be factored

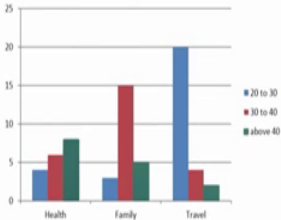


Now, let us try to look at a few simple questions customers were asked to give preferences for color and shape of a product, two team is were created each to determine the color and shape of the product. Is it necessary to check if the two variables are associated, yes it is actually necessary to check it will be actually good if there is no association which means the teams can work independently otherwise the association has to be taken into account in the combination of color and shape that the company finally, chooses to have for the product.


(Refer Slide Time: 20:36)

Question 2

A survey was conducted to understand the reasons for absence of students in a research university. Three main reasons were identified and there were three broad categories of students. In which group, medical reasons dominated? Are the variables associated?



	Health	Family	Travel
20 to 30	4	3	20
30 to 40	6	15	4
above 40	8	5	2



Question number 2, survey was conducted to understand the reason for absence of students in a university the main reasons were identified and there were 3 broad categories of students in which group the medical reasons dominated are they associated.

So, the 3 groups are age groups of 20 to 30, 30 to 40 and above 40 and this is the contingency table. So, we can see from this table that the group of 30 to 40 have a higher family reason and whereas, above 40 had a higher health or medical reason it is possible to find out the association using either chi square or Cramer's V, Cramer's V being a method that also uses chi square we can do that.

(Refer Slide Time: 21:24)




A survey was conducted in a supermarket where 2 variables were considered; let us assume it is a 24 by 7 supermarket. So, 8 am to 8 pm purchase and 8 pm to 8 am purchase and 2 types of customers family and single, would you expect association in the data yes there we expect association in the data and we would expect that more families would come during the 8 am 8 pm and perhaps there will more single people would come during the 8 pm and 8 am.

(Refer Slide Time: 21:57)

Question 4

A survey indicated that the most popular colour for all cars is white. Should a dealer in cars stock all items in white?

Check association between types of buyers and colour.



Survey indicated that the most popular color for all cars is white, should a dealer stock all items in white, now we need to check an association between the types of buyers and color. So, we might have a situation where as in a certain type of buyers white may not be preferred or there could be lesser association and then we have to find out and perhaps the dealer has to stock other colors of cars as well.

(Refer Slide Time: 22:25)


Question 5

Find Cramers V for the following data?

	Red	Blue	White	
More than 30 lakh	20	30	40	90
Between 15 – 30 lakh	10	15	20	45
Less than 15 lakh	40	60	80	180
	70	105	140	315

	Red	Blue	White
More than 30 lakh	$= \frac{90 \times 70}{315} = 20$		
Between 15 – 30 lakh			
Less than 15 lakh			

Cramers V = 0



Now, we looked at last question find the Cramer's V for the following data, let us assume that red blue and white are for example, 3 colors of cars and the other 3 here are people with income more than 30 lakhs less than 15 lakhs between 15 and 30 lakhs and so on.

So, in order to find the Cramer's V or chi square we first try to find the proportion. So, we answer this question if out of 315 70 people have a red color car, now how much out of 90 would have a red color car. So, the proportion becomes 70 by 315 into 90 which happens to be 20. So, if we calculate the remaining numbers which you can do now using the similar formula you would realize that the values are 20 30 40 10 15 20 40 60 80 and the same table repeats when we actually calculate based on the proportions from which the differences will be 0 chi square will be 0 and Cramer's V will also be 0. So, with this we come to the end of this lecture and in the next lecture we would study association between numerical variables.