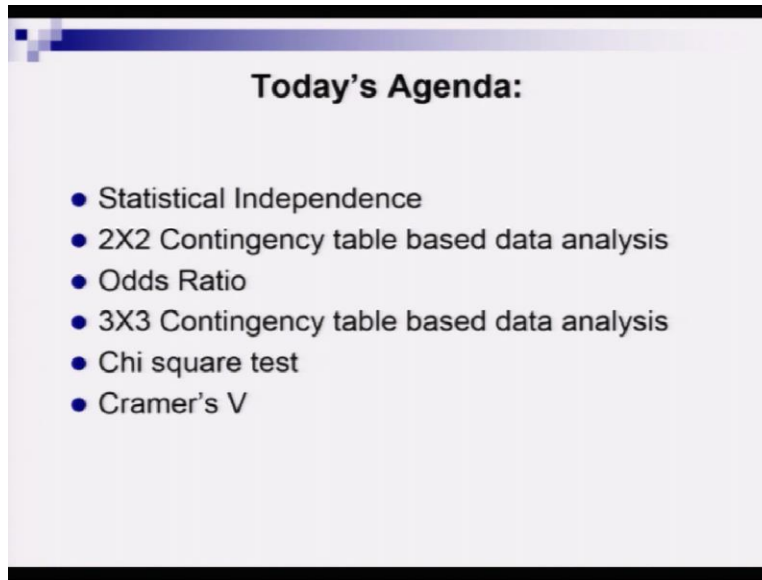


Applied Statistics and Econometrics
Professor Deep Mukherjee
Department of Economic Sciences
Indian Institute of Technology, Kanpur
Lecture-15

Relationship Between Qualitative Variables

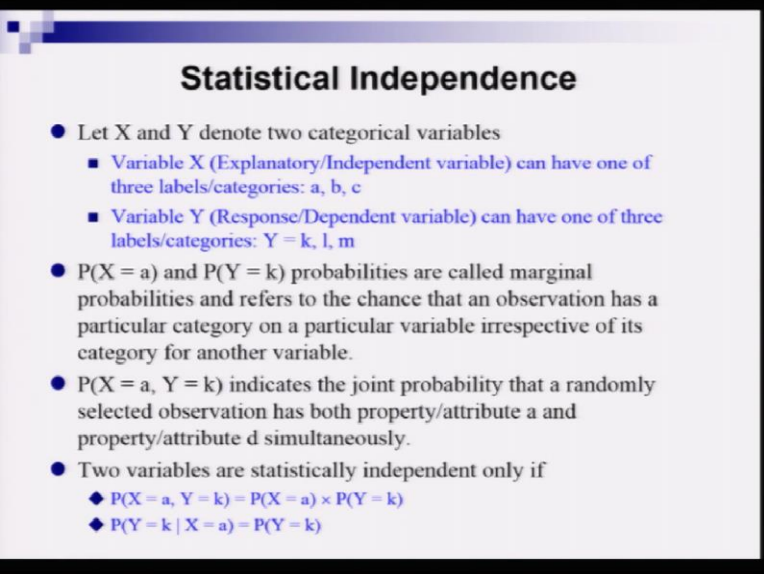
(Refer Slide Time: 00:59)



Hello, friends. Welcome back to the lecture series on Applied Statistics and Econometrics. So, we are almost done with our hypothesis testing discussion. From today's lecture, I am going to start discussion on a new topic or module you can say, and that is basically very interesting in practical purposes and that is called the relationship between 2 variables.

So, here in this lecture, I am going to cover the relationship between 2 qualitative or categorical variables. And in the next lecture, I plan to cover how to model the relationship between 2 quantitative variables. So, let us first look at today's agenda items. So, we will briefly introduce the concept of statistical independence. We are going to discuss in about general contingency table and then, from there, we will study the case of chi-square test. Now, chi-square we have discussed previously, but we are going to revisit that concept again, in a far more general setup. And then, we finally, we will end today's discussion by defining Cramér's V.

(Refer Slide Time: 1:30)



Statistical Independence

- Let X and Y denote two categorical variables
 - Variable X (Explanatory/Independent variable) can have one of three labels/categories: a, b, c
 - Variable Y (Response/Dependent variable) can have one of three labels/categories: Y = k, l, m
- $P(X = a)$ and $P(Y = k)$ probabilities are called marginal probabilities and refers to the chance that an observation has a particular category on a particular variable irrespective of its category for another variable.
- $P(X = a, Y = k)$ indicates the joint probability that a randomly selected observation has both property/attribute a and property/attribute d simultaneously.
- Two variables are statistically independent only if
 - ◆ $P(X = a, Y = k) = P(X = a) \times P(Y = k)$
 - ◆ $P(Y = k | X = a) = P(Y = k)$

So here, we are going to discuss the statistical independence in the context of qualitative variables. So, let us start with denoting 2 categorical variables by x and yes; and x is my explanatory or independent variable. And that can have 1 of the 3 levels or category say ABC, it can have less than 3 also like 2 or it can have more than 3 also. But for simplicity's sake, let us assume that we start with a variable with 3 levels.

Then the variable y is my response variable or the dependent variable. So, by saying that I say that there is some kind of dependence between x and y. So, y actually depends on x that way and this variable also can have 1 of the 3 levels or categories namely k, l and m. So, now, we are going to introduce the concept of statistical independence through probability.

So here, let me define 2 probability terms, probability of capital X taking a level a and probability of random variable y taking a level k. And these probabilities are the marginal probabilities, because this refers to the chances that an observation in contingency table or in our sample has a particular category on a particular variable, say in this case, x equal to a that is irrespective of the category for the another variable.

So, suppose you observe a particular observation number 10 in the data set and for that observation, the, you observe that X has taken value a, but that taking value a is not at all

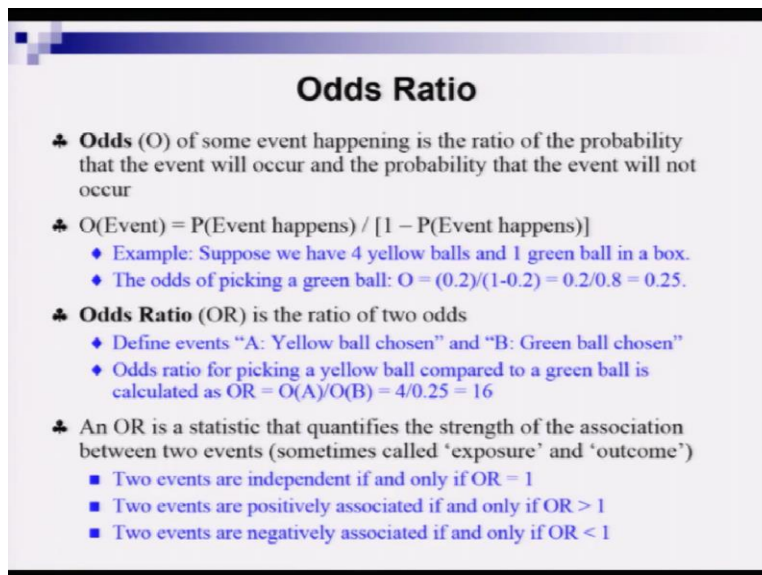
dependent on whatsoever value or level that observation has taken for the other random variable y . So that is why they are truly the marginal probabilities.

So, now we are going to introduce the concept of joint probabilities. So, probability x equal to a and y equal to k actually denotes the joint probability and it means that a randomly selected observation has both property or attribute a , property attribute b , simultaneously. So, here, it is a typo here. So, it should be k and not d .

So now, based on these 2 definitions that we have introduced on probability and we have also talked about conditional probability before, based on these probability concepts, now, we are in a position to define statistical independence. And we provide 2 definitions. 1 is basically in terms of joint and marginal probabilities. And the other 1 is in terms of marginal probability and conditional probability.

Now, we are going to look at a very interesting concept and it has got practical applications in real life and that is the concept of odds ratio. So that is actually pretty much linked with 2 cross 2 contingency table and I would like to start the discussion on odds ratio by saying this that if you are dealing with the higher order contingency table, then the odds ratio calculation may not be possible. So, please remember that whatever we are going to discuss here is strictly valid for 2 by 2 contingency table.

(Refer Slide Time: 05:17)



Odds Ratio

- ♣ **Odds (O)** of some event happening is the ratio of the probability that the event will occur and the probability that the event will not occur
- ♣ $O(\text{Event}) = P(\text{Event happens}) / [1 - P(\text{Event happens})]$
 - ◆ Example: Suppose we have 4 yellow balls and 1 green ball in a box.
 - ◆ The odds of picking a green ball: $O = (0.2)/(1-0.2) = 0.2/0.8 = 0.25$.
- ♣ **Odds Ratio (OR)** is the ratio of two odds
 - ◆ Define events "A: Yellow ball chosen" and "B: Green ball chosen"
 - ◆ Odds ratio for picking a yellow ball compared to a green ball is calculated as $OR = O(A)/O(B) = 4/0.25 = 16$
- ♣ An OR is a statistic that quantifies the strength of the association between two events (sometimes called 'exposure' and 'outcome')
 - Two events are independent if and only if $OR = 1$
 - Two events are positively associated if and only if $OR > 1$
 - Two events are negatively associated if and only if $OR < 1$

So, to study odds ratio, we have to first study what is odds. So, odds are basically related to the probability of an event, but they are not exactly the probability of a particular event happening. So, odds is defined as the ratio of the probability that that particular event will occur and the probability that the event will not occur. This formula from there, you can actually get back the probability of a happening very easily, if you switch terms. So, in that case, probability of event a happening can be calculated as odds divided by 1 plus odds.

So, now, let us explain the idea of odds with a very simple example. And this example we are going to continue with when we are going to introduce the concept of odds ratio. So, let us assume that we have a box and there we have some balls and we see that there are 5 balls, 4 yellow and 1 green. So, now we are going to close our eyes. And we are going to draw a ball randomly from that box. So, this is basically the experiment. And then, in this context, let us take examples of odds and odds ratio.

So, in this particular context, what is the odds of picking a green ball? So, you see the probability of picking a green ball from that box containing 4 yellow and 1 green is basically 1 over 5 and that is basically 0.2. And when we are talking about calculating odds, so, it is not only probability of picking up a green ball from the box, we also have to calculate the probability of yellow ball picked up from the box.

So, in that case, of course, it will be 1 minus 0.2 or, as there are 4 yellow balls in the box. So, the probability that a yellow ball will be taken randomly is 0.8. So now, you have to take the ratio of 0.2 and 0.8. And then, you will get 0.25 as the odds of picking a green ball, which is different from the probability of picking a green ball which is 0.2.

Now, we are going to talk about odds ratio and odds ratio, OR is defined as the ratio of 2 odds. So, in this particular example of box containing some balls, there are 2 likely events either a yellow ball will be chosen or a green ball will be chosen. So, let us define an event A, which says the yellow ball has been chosen. And let us define another event B, where we can say that it is a green ball which was chosen. So, now, I am interested to calculate the odds ratio for picking a yellow ball compared to a green ball.

So basically, we are interested in OA divided by OB. So, this is the ratio of 2 odds for events A and B, respectively. So, I am not showing you how to get the OA, which is basically the odds for event A, you can calculate it yourself as an exercise following the previous example of odds of picking of a green ball. And that is 4 and that has to be now divided by the odds of picking a green ball because that is my event B. So, I divide 4 by 0.25 and then, I get 16.

So, how can I interpret odds ratio? Odds ratio has very nice interpretations, but it depends on specific context. So, let us save the interpretation of odds ratio for a moment. Now, what can we say more by looking at the odds ratio that we have calculated? Well, we can talk about the degree of association between 2 qualitative variables by looking at the odds ratio and if odds ratio takes value 1, then that means that 2 events are equally likely, and basically, they are independent of each other.

And if odds ratio is higher than 1, then basically we can say that 2 events are positively associated or 2 variables are positively associated. And if we observe odds ratio, being less than 1 then we can say that these 2 variables are negatively associated.

(Refer Slide Time: 09:57)

2×2 Contingency Table

- Two-way frequency distribution a.k.a. bivariate frequency table
- Let X and Y denote two nominal/categorical variables
 - X (Explanatory/Independent variable) can have one of two values/labels: X = 1 or X = 2
 - Y (Response/Dependent variable) can have one of two values/labels: Y = 1 or Y = 2
- Let n_{ij} denotes the count of observations in a cell in a table

X / Y	Y=1	Y=2	Total
X=1	n_{11}	n_{12}	n_{1+}
X=2	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

Probability examples:

- *Conditional probability:*
 $P(Y = 1 | X = 1)$
 $= n_{11} / (n_{11} + n_{12})$
- *Joint probability:*
 $P(Y = 1 \text{ and } X = 1)$
 $= n_{11} / (n_{11} + n_{12} + n_{21} + n_{22})$

Now, we are going to look at the 2 by 2 contingency table. It is the simplest possible tabular format of representation of qualitative data. And we have already seen this twice in this course. Once at the very beginning and then, 1 at the time of the chi-square independence test that I studied maybe 2 lectures before.

But here, I am going to now link this odds ratio concept with contingency tables that we have not discussed before. So, if you remember from our previous discussions, 2 by 2 contingency table is basically a two-way frequency distribution. It is also called a bivariate frequency table. And as again, let us assume that we are dealing with 2 nominal or categorical variables x and y.

But here, is the difference; here, we are assuming that we have explanatory variable x which takes only 2 values or levels and let me assume that they are 1 and 2. This is just for simplification purpose. And then my yes, the dependent variable or the response variable can have 1 of the 2 values or levels and again, I am simplifying by assuming yes can take values 1 or 2. Now, n_{ij} denote the count of observations in a cell in the contingency table. So, here I am showing you the contingency table. So, now I am going to show you 2 examples, how to compute conditional probability and joint probability from a 2 by 2 contingency table.

So now, come back to the table. Here, I am showing you that the case of conditional probability for cell 1, which has the frequency n_{11} , and what is the probability that y takes value 1 given that x is 1. So, for that, what we have to do, we have to basically in the numerator, we will have n_{11} because that is the frequency which satisfies both x equal to 1 and y equal to 1. And then, basically what I have to do, I have to divide it by the sum of the numbers that you see in 2 columns for x equal to 1.

So, as x is given 1, so then you have to sum all the frequencies, which is satisfying that criteria that x equal to 1. So, x equal to 1 case can have 2 different values of y , right? y equal to 1 and y equal to 2, n_{11} number of observations taking value y equal to 1 and n_{12} number of observations taking value y equal to 2. So, ideally, you need to sum them up and then the ratio of n_{11} and the sum will give you the conditional probability.

Now, we are going to look at the joint probability again for the first cell and that is basically the probability statement, probability y equals to 1 and x equal to 1. Note that, I am not saying that I am computing the probability of 1 for the case when x takes value equal to 1. So, I am talking about the joint occurrence of y equal to 1 and x equal to 1. So here, what will happen, so here, you see that n_{11} will be the numerator again and in the denominator now, you have to sum all four frequencies n_{11} , n_{12} , n_{21} and n_{22} , why is this so? Because we are talking about the joint occurrence of y equal to 1 and x equal to 1.

So, note that, how many ways actually y equal to 1 can happen. So, y equal to 1 can happen in 2 ways, when x equal to 1 and when x equal to 2 and there are 2 frequencies there in n_{11} and n_{21} . And then how x equal to 1 can happen, x equal to 1 can happen for 2 cases here. So that is n_{11} and then n_{12} . So, you have to basically divide the frequency that you observed in the cell x equal to 1, y equal to 1 and divide that by the total number of observations in the sample, which is n .

(Refer Slide Time: 14:37)

Example: Odds Ratio in 2×2 Crosstab

X / Y	Y=1	Y=2	Total
X=1	n_{11}	n_{12}	n_{1+}
X=2	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

	Pass	Fail	Total
Trained	16	7	23
Not	8	11	19
Total	24	18	42

- ♣ A simple formula for OR = $(n_{11} \times n_{22}) / (n_{12} \times n_{21})$
- ♣ Example: We have two samples of job-seeking candidates. Sample one consists of 23 students from a coaching/training program. Out of them 16 are able to pass a selection examination but 7 are unable. The other sample comprises of 19 persons those that didn't take the training. Out of them, only 8 passed.
 - O(Someone passing if s/he took training) = $16/7 = 2.29$
 - O(Someone passing without training) = $8/11 = 0.73$
 - OR(Passing after taking coaching) = $(16 \times 11) / (7 \times 8) = 3.142$

Now, we are going to talk about the odds ratio calculation through the 2 by 2 contingency table by looking at a very simple hypothetical example. So, before I go to the example, let me know have a depiction of the formula that a very simple formula for OR calculation and that is basically given at the first bullet point and that is n_{11} times n_{22} , these are the frequencies in cells and that is divided by n_{12} times n_{21} .

So basically, it is very simple to remember. So, you start from the northwest corner of the table or the northwest frequency cell and then, you come to the southeast cell and that you have to come diagonally, you need to multiply these 2 frequencies. And then again, you have to move up, you have to go northeast corner of the table to the cell and then, you have to come down to southwest of the table to the frequency cell and you need to multiply. And if you take ratio, then you get the OR or the odds ratio.

So, note that this is very interesting, when I first introduced the concept of odds ratio, I said that it is a ratio of 2 odds, and then to calculate the odds ratio from a theoretical point of view, we have to first get to know the probability of an event happening and then, you have to calculate the probability of that event not happening.

And then, similarly, you have to do that for the another event, B say. So, you have done it once for event A and then you have done it for event B and then you calculate the ratio

and that is your odds ratio. But when you are dealing with 2 cross 2 simple contingency table, actually you can skip that probability computation and you can go for direct computation of odds ratio from the frequency itself, that is what it was shown through that formula.

Now, let us look at a very simple example. So, we all know that these days parents send their sons and daughters to different coaching centers, so that their chance of passing particular entrance exam for either entering a higher education institute or getting a government job or other types of jobs increases. So, basically the prospects of peering exams increase.

So here, we have 2 samples of job seeking candidates. And sample 1 consists of 23 students from a coaching or training program and these 23 students actually have attended the training for cracking examinations and out of these 23, 16 of them are able to pass particular examination, but 7 could not.

Now, you also get another sample which comprises of 19 persons and they actually somehow did not attend a coaching or training program. So, they made all the preparations themselves for the examination. So, out of these 19, now only 8 could pass a particular selection examination and 11 failed.

So, this is basically the context and if this is the story I have, how can I represent this hypothetical story in terms of a 2 cross 2 contingency table. So, now, let us look at the table at the right-hand side; and then, here you see along the rows, I am going to measure the explanatory variable, which is attending a particular training or coaching program or not?

And along the columns, I am going to measure the response or the outcome variable and that is basically pass or fail in a particular selection examination. So now, in 4 cells I have placed the frequencies that I have got from the example. And then, I also have shown you the totals.

Now, I am going to apply the magic formula that is in the first bullet of the slide to these numbers and then let us get the odds ratio. But it is not a bad idea if we use these

numbers in the cells, also to calculate the odds for different events. Because odds with which we started I think we should have some example for that concept as well.

So, I define my first odd as the odd of someone passing, if he or she took a particular training program. And you see here that out of those 23, 16 passed and 7 failed. So, the odds has to be 16 divided by 7 and that gives you the number 2.29. So, you can interpret this number as their odds of passing the examination is 2.29 times higher for a student if the student has taken a training.

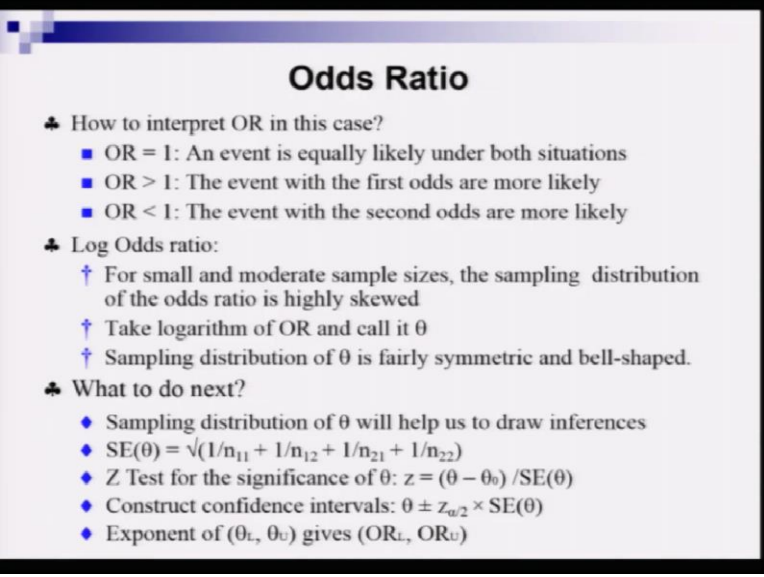
And how do you then calculate the odds for someone passing without training? So here you can see that the numbers are 8, 11 and 19, 19 is total. So only 8 passed and 11 failed. So, you have to take ratio of 8 and 11 and then, you get 0.73.

And how do you interpret this number? So, you can say that if a student or candidate has not taking a coaching or training program, then he or she is 0.73 times likely to pass compared to fail? And now how do you define the odds ratio overall? So here, you define the odds ratio as passing after taking the coaching program.

And here, you apply the formula from the sales that you are seeing here in the left hand side table and the numbers are shown for this case study knowing the right hand side tables, so you may get the values of n_{11} and n_{22} and they are basically 16 and 11. And then, you multiply and then divide by these 2 numbers n_{12} and n_{21} , which are respectively 7 and 8 from our table. And finally, you get the value 3.142 as the odds ratio.

Now note that, it is very interesting, if you now take the ratio of 2 odds that we have calculated beforehand, which are 2.29 and 0.73. So, that is odds of the event called someone passing if the candidate has taken a training. And the second event was that someone passing without taking a training. So, if you take the ratio of 2.29 and 0.73, these 2 odds numbers you actually get 3.142, okay?

(Refer Slide Time: 22:13)



Odds Ratio

- ♣ How to interpret OR in this case?
 - OR = 1: An event is equally likely under both situations
 - OR > 1: The event with the first odds are more likely
 - OR < 1: The event with the second odds are more likely
- ♣ Log Odds ratio:
 - † For small and moderate sample sizes, the sampling distribution of the odds ratio is highly skewed
 - † Take logarithm of OR and call it θ
 - † Sampling distribution of θ is fairly symmetric and bell-shaped.
- ♣ What to do next?
 - ◆ Sampling distribution of θ will help us to draw inferences
 - ◆ $SE(\theta) = \sqrt{1/n_{11} + 1/n_{12} + 1/n_{21} + 1/n_{22}}$
 - ◆ Z Test for the significance of θ : $z = (\theta - \theta_0) / SE(\theta)$
 - ◆ Construct confidence intervals: $\theta \pm z_{\alpha/2} \times SE(\theta)$
 - ◆ Exponent of (θ_L, θ_U) gives (OR_L, OR_U)

So, now, again, we are going to look at the interpretation of the odds ratio. So, I have told you previously that odds ratio can be seen as a measurement of degree of association between 2 qualitative variables. So, in this context, how do I interpret the odds ratio? So here, if the odds ratio is taking value 1 then that means that the event is equally likely under both situations.

Now, what do I mean by the event is equally likely under both situations? So, it is basically the event of passing the exam and what is these both situations? So, situations here are either you have taken the training or you have not taken the training. So, these two are 2 different situations.

Now, the second sub-bullet of the first bullet point says that if the OR score is greater than 1 then event with the first odds are more likely to happen. And that is basically the event we are talking about that is the first event that the student has passed a particular selection exam after taking the training program. And if the OR is less than 1, then the event with the second words are more likely. So here, the second odds basically passing the exam, but without attending the training program.

So here, you see that now, we got a OR score of 3 point something from the previous slide. So here, as the OR value is greater than 1, we can say that these 2 qualitative

variables passing a particular selection examination and attending a training or coaching program, they are positively related. So that means that if the student actually has taken the training program before appearing for the examination then the odds are higher for him or her to pass the exam.

Now, we are going to talk about the concept called log odds ratio. Now, why we have to talk about the log odds ratio because sometimes it is purely an empirical issue that sometimes odds ratios, sampling distribution becomes very skewed. And I hope that when I say that sampling distribution of odds ratio, you will understand what I am trying to say, because you have calculate the odds ratio from 1 particular sample after tabulating the numbers in the 2 by 2 contingency table.

But what if, if you take another sample? So of course, you will get into another odds ratio. So then, if you plot all these odds ratios then you are going to get the sampling distribution of that, right? So now, empirically, what we have seen that the sampling distribution of the odds ratios can be very skewed at times and then, it is not a very good feature, when you are trying to conduct hypothesis testing and estimation.

So basically, statisticians propose that why do not we take logs? The first step is to calculate the standard error of log odds ratio θ and how do you calculate that? So, there is a formula, I am showing. So, basically you have to take the inverse of the sale frequencies, you need to sum these inverse sale frequencies and then, the sum the square root of the sum should be taken and that will give you the standard error of θ .

So here, the second sub-bullet point under the last bullet point in the slide is showing you the formula for that. And as we are assuming that our sampling distribution of θ is more or less symmetric bell shaped and normal approximation, we can assume so, why do not we go for as a z test? Yes, we can actually conduct a z test for the significance of θ .

So, you have to calculate or define the z statistic for that. We assume that the unknown population parameter value of θ is θ_0 . So, in that case, we define z as θ

minus θ_0 divided by standard error of θ . This is very common to what we have done previously when we conducted z test.

So, after we have defined the z test, we can also construct the confidence intervals for these parameter θ . And that is basically $\theta \pm z$ for the significance level α by 2 multiplied by standard error of θ . So, $z_{\alpha/2}$ is basically also not unknown to you, it is basically the critical value from the standard normal table for the level of significance α , as we are doing a 2-sided thing, so, that is you need to divide it by 2.

And then, once you get the number from the statistical table, you multiply that with the standard error that you can calculate using the formula in this slide only. And then, that number, that product has to be added and subtracted to the sample statistic value, which is θ here. And then, you get your confidence interval.

So, once the confidence interval is constructed for θ , note θ is basically not the entity with which we are going to work. We are actually finally interested in the odds ratio. So how do I get the confidence interval for odds ratio? Simple, because you have taken the log to get the log odds ratio, so you have to reverse the process. So, basically, the limits of the confidence interval now, you will have to take the exponent of those numbers and then, these new numbers will give you the confidence limits for the odds ratio.

So, that is what I am showing here, that if you have θ_L and θ_U the limits of the confidence interval, then you take exponents of these 2 numbers and you get ORL and ORU respectively. So, now, let me emphasize on 1 limitation of this 2 by 2 tables. And that is why we can go for generalization First of all, you may be interested in 2 qualitative variables, which will have more than 2 levels or categories or attributes that you are interested in. And why then restrict yourself to only 2 particular levels. And the second is that well, the odds ratio you cannot compute in 2 cross 3 or 3 cross 2 or 3 cross 3 or higher dimension contingency tables.

So, how do you then measure the degree of association shape between 2 qualitative variables? So then actually, we have to go for something general in nature and for that we have to introduce a new concept of degree of association between 2 quality variables and that is called Cramér's V.

But to see how Cramér's V is computed, let us tell the story from a very general point of view. So, we are not going to show you an illustration of 3 cross 3 or 4 cross 4, because the solution or the method we are going to talk about is far more general. So, let us now look at a m cross p contingency table.

(Refer Slide Time: 29:32)

Contingency Table and Independence

- We have n observations on two categorical variables X and Y
 - X can have one of p categories/labels and represented in rows
 - Y can have one of m categories/labels and represented in rows
 - Let f_{ij} denotes the frequency corresponding to i -th label of X and j -th label of Y
- The χ^2 statistic compares an observed frequency table produced by the sample with a hypothetical frequency table that would occur if the variables are indeed statistically independent (in the population)

$$\chi^2 = \sum_{i=1}^p \sum_{j=1}^m \frac{(f_{ij} - e_{ij})^2}{e_{ij}} = \frac{(f_{11} - e_{11})^2}{e_{11}} + \frac{(f_{12} - e_{12})^2}{e_{12}} + \dots + \frac{(f_{1m} - e_{1m})^2}{e_{1m}}$$

$$+ \frac{(f_{21} - e_{21})^2}{e_{21}} + \frac{(f_{22} - e_{22})^2}{e_{22}} + \dots + \frac{(f_{2m} - e_{2m})^2}{e_{2m}}$$

$$+ \dots + \frac{(f_{p1} - e_{p1})^2}{e_{p1}} + \frac{(f_{p2} - e_{p2})^2}{e_{p2}} + \dots + \frac{(f_{pm} - e_{pm})^2}{e_{pm}}$$

So now, let us assume that like previously we have a n observations So, sample size is n and we have these observations on 2 categorical variables x and y. But note that there is this difference here. So, x can now have 1 of the p categories or levels and these p levels or categories are represented in p rows. And then the y, the response variable can also have 1 of the m categories or levels, and they are represented in the different columns.

So here, it is a typo here, so we will correct this to columns. So, the second sub-bullet point in the first bullet point, you make this note that y is represented through the columns and not the rows.

Let us now assume that f_{ij} denotes the frequency corresponding to the i -th level of x and the j -th level of y . So, if you remember, when we introduced chi-square statistic, for the test of independence previously, we have said that to compute the chi-square statistic, you have to first actually compute the expected frequency.

Then you have to basically take the difference of the actual or observed frequency and the expected frequency then, square it and then, basically you have to divide it by again the expected frequency. And then you sum over all cells. And that is how you get your chi-square statistic.

So, but that was the 2 cross 2 case when we studied it last time. So, what will happen if we have this generalized matrix form of p cross m dimension? So that is what I am showing you here in the clumsy expression at the bottom of the slide. So, chi-square is defined as double sum and i is basically giving me the number of rows from 1 to p and small j is giving me the number of columns from 1 to m .

And then as I have explained you a short while ago that we have to construct the difference of the expected frequency and the observed frequency. So here, in this formula, observed frequency is f_{ij} . So, that is basically you observe in the contingency table that you have constructed from your data.

Then you have to calculate the expected frequencies and that is basically given by e_{ij} . e_{ij} is basically the expected frequency for the i comma j -th cell. So, i here refers to the row, and j here refers to the column. So, then the difference has to be squared and then, you have to divide this by the expected frequency for that particular cell and then, you need to sum over different cells.

So here, the basic philosophy is very simple. So the constructed or calculated chi-square statistic, basically does a very simple thing. It compares the observed frequency table produced by the sample with a hypothetical frequency table that would occur if the variables are indeed statistically independent in the population. So, what do I mean by that? How do you then compute the complicated expected frequencies? Well, it is not very complicated actually. If you remember the previous discussion, when we had for the

chi-square test of independence, I have given you a formula and that was basically for a particular cell, the expected frequency is given by the row total multiplied by the column total divided by the grand total.

So now, let us look at that same formula here in this general case and let us see how we can calculate the expected frequency for various sales. So, I will start here with the very first component in the sum expression. So, that is f_{11} minus e_{11} square divided by e_{11} . So, what does it mean? So f_{11} is basically the observed frequency that your data tells you in the x equal to 1 and y equal to 1 cell. And then, e_{11} is basically the expected frequency that you need to now calculate. F_{11} you cannot calculate, it is given to you, e_{11} you have to calculate.

Now, how would you calculate? So basically, what you do? So here, you see the row references 1, column reference is also 1. So, what you have to do? Here, you have to concentrate on row number 1 in the table and then, you need to sum the frequencies across columns.

But you have to stay on the same row and that will give you the row total. And then, you have to now concentrate on the column. And then, now, you need to move down the rows and then keep on adding the frequencies.

So that will finally give you the column total. So, you need to multiply that first sum, which is the row total. And the second sum, which is the column total. And then you divide that product by the total number of observations in your data set, which is small m , which is the grand total. And then basically, that is your expected frequency. So, you take the difference, square it and then you finally divide that in term by the expected frequency itself. That is what the first component in this sum expression tells us. Similarly, you can follow the same logic for the other components that you are seeing here in this clumsy expression.

(Refer Slide Time: 35:44)

Chi-squared Test for Independence

- How to calculate expected frequency? $e_{ij} = n \times p_{i\cdot} \times p_{\cdot j}$
- Simply, $e_{ij} = (\text{row } i \text{ total} \times \text{column } j \text{ total})/n$
- $\chi^2 = 0$ implies statistical independence which means no association
- To conduct a test, ideally $e_{ij} \geq 4$ for every i and j
- H_0 : Row variable X **is independent** of column variable Y
- H_1 : Row variable X **is not independent** of column variable Y
- If χ^2 is past a critical value then there is reason to reject the null hypothesis that the two variables are independent
- Find critical value $\chi^2_{(p-1)(m-1),\alpha}$ from Table such that $P(\chi^2_{(p-1)(m-1)} > \chi^2_{(p-1)(m-1),\alpha}) = \alpha$
- Decision rule:

Reject H_0 if $\chi^2 > \chi^2_{(p-1)(m-1),\alpha}$
Do not reject H_0 if $\chi^2 \leq \chi^2_{(p-1)(m-1),\alpha}$

So here in this slide, I am going to repeat the same thing again for your help. So here, I am showing you a formula. Now, you get your chi-square value and from there actually, how would you interpret the matter? How can you decide whether 2 qualitative variables are indeed associated or not? So, if you by chance get a chi-square value of 0, then that means that the variables are actually independent, they are not associated with each other.

But here, there is a note of caution that you must remember. And it is that fact that you cannot make such a strong decision based on one particular sample, because there could be sampling fluctuation. So, what if you get another sample from the population and then, that will give you a chi-square value of say, 0.19. Or it can also give you a chi-square value of 0.73, then how would you actually compare these 3 chi-square numbers because they are telling you different values, but these are from the 3 different samples.

So how would you conclude whether the 2 variables are independent or not. So, for that, actually, you have to conduct a hypothesis testing. And that is what the Pearson's chi-square test does for you. So, here, I am going to describe, again, Pearson chi-square test of independence. And just again, another note of caution, to conduct Pearson's test ideally, your e_{ij} should be greater than or equal to 4 for every i and j otherwise, there could be some abnormalities.

So, that is basically a note of caution. But suppose our cells satisfy these criteria. So, in that case, we will continue with the formation of null and alternative hypothesis. So, here null hypothesis will be that my row variable x is independent of column variable y and my alternative hypothesis would be that the row variable x is not independent of the column variable y .

So, basically, H_0 says that there is no association and H_1 says that there is association between these 2 qualitative variables. So, now you have to then fix a particular level of significance say α and then, you have to find the critical value, so that you can actually compare the critical values and the test statistic value and then, take a decision.

So, if the chi-square is higher than critical value, then there is reason to reject the null hypothesis that 2 variables are independent. So, that is basically the philosophy of the rejection that we have followed, if we are following the traditional approach or traditional method that we have done several times in the past in this course only.

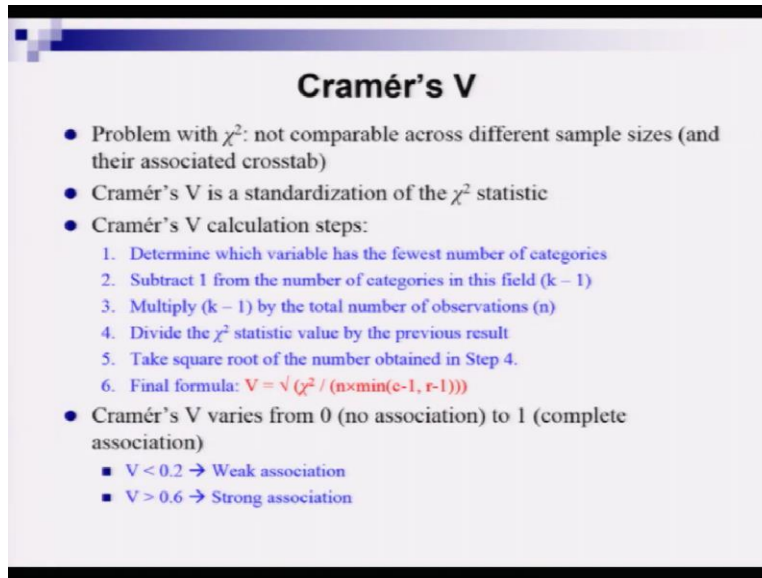
So now, how to find the critical values to be specific? So here, find the critical value now, you have to basically consult the chi-square table, right. So, now chi-square test statistic, will follow the chi-square distribution with degrees of freedom $p - 1$ times, $m - 1$. So, that is basically the number of rows minus 1 times number of columns minus 1. And for an α , you consult the table and you get some value. And then, basically, that is basically your critical value, right.

So, here, the decision rule says that you can reject your H_0 if the calculated or observed chi-square statistic value is higher than the chi-square tabulated value that you find from the table or in other words, the critical value. And you do not reject your null hypothesis if chi-square actually is less than or equal to the tabulated value or the critical value that you find from the table.

So, now we are at the fag end of the discussion on contingency table and degree of association between 2 qualitative variables. So, I said that I am finally going to talk about a measure Cramér's V , which can be used for qualitative variables involving more than 2

levels or categories. And here you go. So, in the last slide, we are going to talk about this Cramér's V only.

(Refer Slide Time: 40:41)



Cramér's V

- Problem with χ^2 : not comparable across different sample sizes (and their associated crosstab)
- Cramér's V is a standardization of the χ^2 statistic
- Cramér's V calculation steps:
 1. Determine which variable has the fewest number of categories
 2. Subtract 1 from the number of categories in this field ($k - 1$)
 3. Multiply ($k - 1$) by the total number of observations (n)
 4. Divide the χ^2 statistic value by the previous result
 5. Take square root of the number obtained in Step 4.
 6. Final formula: $V = \sqrt{\chi^2 / (n \times \min(c-1, r-1))}$
- Cramér's V varies from 0 (no association) to 1 (complete association)
 - $V < 0.2 \rightarrow$ Weak association
 - $V > 0.6 \rightarrow$ Strong association

So, note that, chi-square test also has got some problem. Now, what is the problem? So, the chi-square test is not standardized, because if you take a data set and then, you compute the chi-square test statistic value and observe some value, you can get another sample and then, calculate another chi-square value these 2 chi-square values are not comparable to each other.

And also suppose, you have 1 researcher who has talked about the same problem, but his x and y has say 2 levels each, but the other researcher has dealt with the same research problem, but that researcher or researcher number 2 is showing you a contingency table, where x and y have more levels say 3 each. So, from these 2 different researchers, you can see 2 different chi-square values computed. But you cannot compare because the dimension of the contingency table is not matching.

So, how can you actually compare chi-square statistic values across different contingency tables? So, we need some standardization. So Cramér's V actually helps you to look at this problem. So, Cramér's V actually is a standardization procedure of the calculated chi-

square statistic, and its calculation involves several steps. So, before I show you the final formula, let me take you through the, or walk you through the steps.

So, in the first step, you have to first determine, which variable has the fewest number of categories. And then you have to subtract 1 from that number of categories in this particular field. And suppose your row variable has fewest number of categories and it has k category, so, you have to deduct 1. So, the point of interest number will be k minus 1. Now, you in the third step have to multiply that k minus 1 to the total number of observations n . And in fourth step, you need to divide the chi-square statistic that you calculated by the number n times k minus 1.

And finally, in step 5, you take the square root of the number opted in step 4. So, in bullet point or sub-bullet point 5 of the main bullet point 3, I am showing you the final formula V . So, note that, the beauty of the Cramér's V is that, this Cramér's V is such a statistic that its values are bounded. So, its values are bounded by 0 and 1. So, minimum possible value is 0 and the maximum possible value is 1. Whereas, chi-square actually has no upper bound, chi-square can only be 0 at the no minimum, but it can take any positive number.

So, that way Cramér's V is a standardized version of chi-square statistic and its value actually helps us to comment on the association between 2 qualitative variables. So, if the value of V is less than 0.2, then you can say that well the association between 2 qualitative variables is not that strong. But if it is, if the V 's value is higher than 0.6 then you can say that well, there is evidence from sample that there is strong association between these 2 qualitative variables.

Okay. So, this is the end of discussion on the degree of association ship and relationship between 2 qualitative variables. So, in the next lecture, we are going to talk about the relationship between 2 quantitative variables. And now, we are going to talk about specifically the concepts of correlation and regression. See you then, thank you.