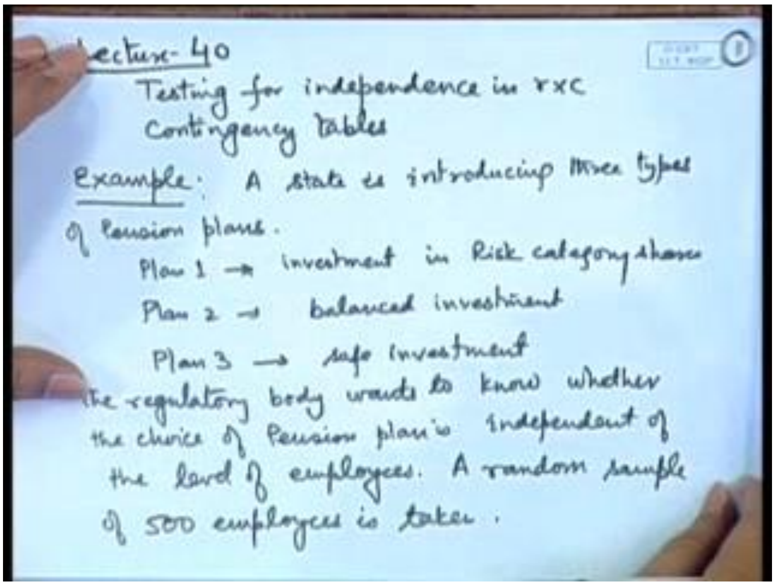
Probability and Statistics
Prof. Somesh Kumar
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture - 78
Testing for Independence in rxc Contingency Table – 1

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

(Refer Slide Time: 00:59)



Lecture- 40
Testing for independence in rxc
Contingency tables

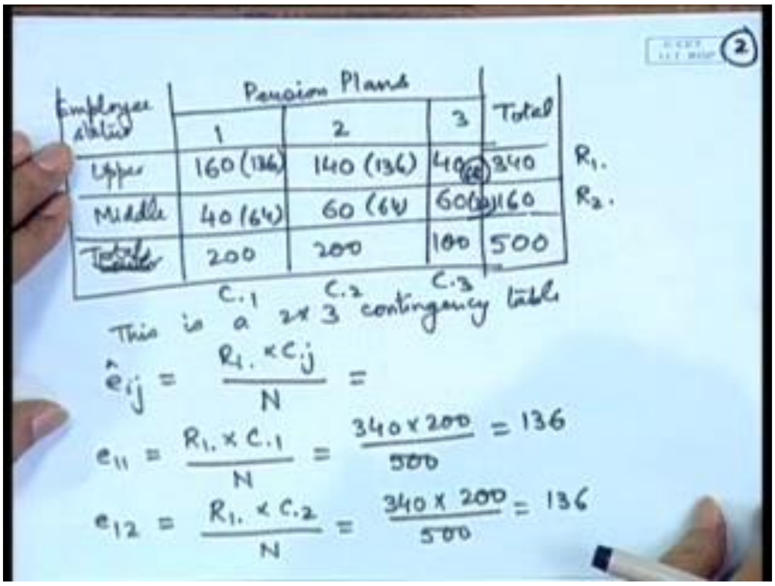
Example: A state is introducing three types of pension plans.
Plan 1 → investment in Risk category shares
Plan 2 → balanced investment
Plan 3 → safe investment
The regulatory body wants to know whether the choice of pension plan is independent of the level of employees. A random sample of 500 employees is taken.

So, this topic we called as testing for independence, testing for independence in r by c contingency tables.

So, yesterday I mention that we may have r categories distributed in the rows and c categories which are represented in the columns, we have observed frequencies o i j's corresponding to i j th cell and on the basis of this we want to test whether the 2 categories are independent. Let me explain this through an example here. So, the problem is posed as follows. So, a state is introducing 3 types of pension plans. So, in the plan 1 the investment of the pension fund will be in risk category shares; in plan 2 this is balanced investment and plan 3 is say for safe investment. Now whether the employee’s preferences are affected by their hierarchical structure in the organization that is what we have to check.

So, the regularity in a regular regulatory body wants to know whether the choice of pension plan is independent of the level of employees. So, a random sample of 500 employees is taken and we observe the following data.

(Refer Slide Time: 04:03)



2

Employee status	Pension Plans				
	1	2	3	Total	
Upper	160 (136)	140 (136)	40 (68)	340	$R_{1.}$
Middle	40 (64)	60 (64	60 (32)	160	$R_{2.}$
Total	200	200	100	500	
	$C_{.1}$	$C_{.2}$	$C_{.3}$		

This is a 2×3 contingency table

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

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

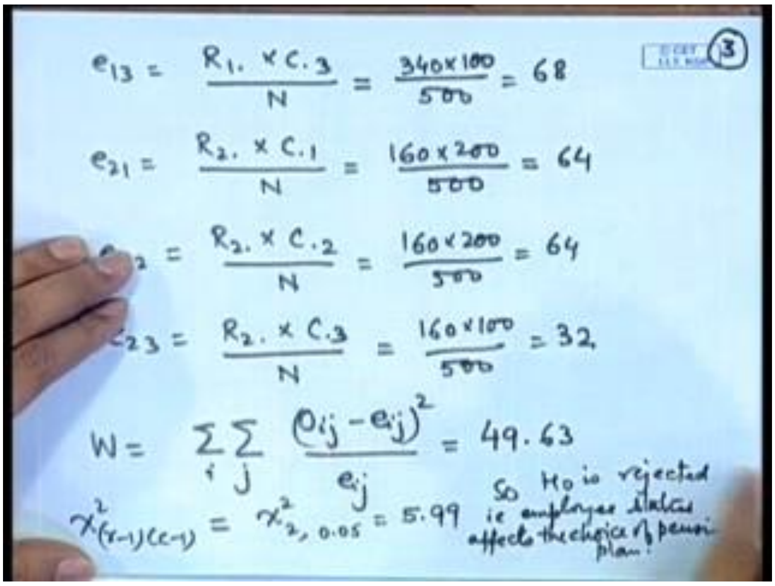
$$e_{12} = \frac{R_{1.} \times C_{.2}}{N} = \frac{340 \times 200}{500} = 136$$

So, the choice of pension plans we represent in the column that is plan 1, plan 2 and plan 3; and here we give employee a status say. So, we have upper level, middle level and say lower level. For the time being let me concentrate only suppose upper level and middle level here and the data of 500 is distributed as 160 upper level employees give preference to pension plan 1; 140 give to pension plan 2 and 40 give to pension plan 3.

In the middle level 40 give preference to pension plan 1, 60 to 2 and 60 to 3. So, if we calculate the row and column totals, they turn out to be 200, 200 and 100 and here it is 340, 160. Now we want to test whether the choice of pension plan is independent of the employee’s status. So, for this, this is a 2 by 3 contingency table, this is a 2 by 3 contingency table. So, the values of the e i j's they are calculated by R i into c dot j divided by n. Now here you see the rho total that is R 1 dot, R 2 dot, this is c dot 1, c dot 2, c dot 3. They are given to us so easily we can calculate say e 1 1. So, e 1 1 will be R 1 dot into c dot 1 by N.

Now, in this particular case it is 340 into 200 divided by 500, this value turns out to be 136. So, we can write these values here, similarly if I want to calculate e 1 2; e 1 2 is R 1 dot into c dot 2, divided by N that is equal to 340 into 200 divided by 500 there is again 136.

(Refer Slide Time: 07:17)



3

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

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

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

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

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

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

So H_0 is rejected ie employee status affects the choice of pension plan.

Similarly, we can calculate e 1 3 that is R 1 dot into c 2 3, divided by N that is equal to 340 into 100 divided by 500, that is equal to 68. So, this value is 68.

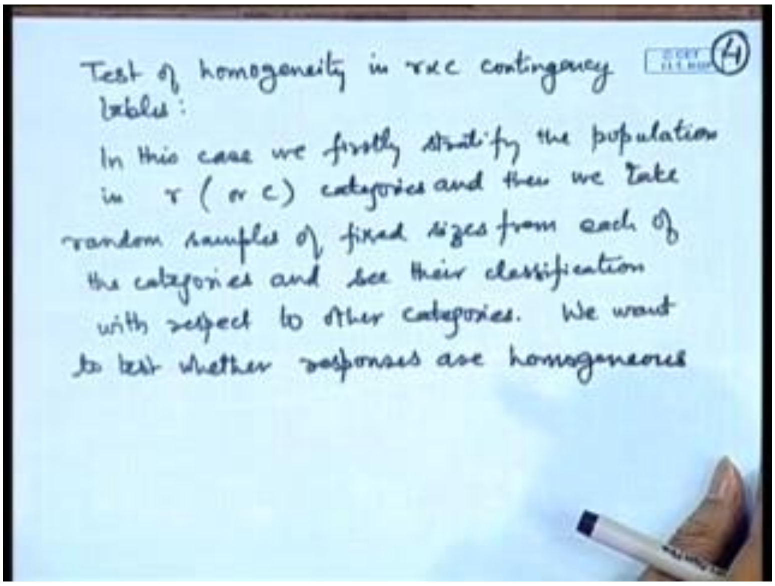
In a similar way we can calculate the value corresponding to e 2 1. So, e 2 1 will be R 2 dot into c dot 1 divided by n that is equal to 160 into 200 divided by 500 that is equal to 64 e 2 2 that will be equal to R 2 dot into c dot 2 by n, that is equal to 160 into 200 divided by 500 that is again 64, and e 2 3 that is equal to R 2 dot into c dot 3 divided N that is equal to 160 into 100 divided by 500 that is equal to 32. So, we can complete this table of e i j's here. So, this is 64, this is 64, this is 32.

Now, our formula for the W is sigma of o i j minus e i j square divided by e i j. So, these differences can be calculated. So, for example, the first term is 160 minus 136 that is 24 square divided by 136, again similar term here this will become 4 square by 136, here it is 28 square by 30 by 68 and like that. So, we can calculate these terms the overall W turns out to be 49.63. Now here the calculated value of the chi square statistic will be on r minus 1, c minus 1 degrees of freedom.

Now, here 2 rows are there and 3 columns are there. So, this becomes chi square 2 and we look at say 0.05 etcetera then this is giving the value 5.99, which is much smaller. Actually we can calculate at a very small level of significance and this value will be still be larger. So, H naught is rejected, what is H naught? H naught is the hypothesis that the rho categories and the column categories are independent, that is H naught is rejected; that means, the employee status affects the choice of pension plan there is another application of the chi square test, in the contingency table we have seen that we took a total sample of size n and then we saw the actual classification in the r c categories.

But sometimes we may fix the data for example, we may take a fix number from upper income group, upper level employees, we may take a fixed sample size from middle strata. So that means, basically we are doing the stratification of the population and then we take the sample and we want to see whether the responses of the different strata are homogeneous.

(Refer Slide Time: 11:32)



4

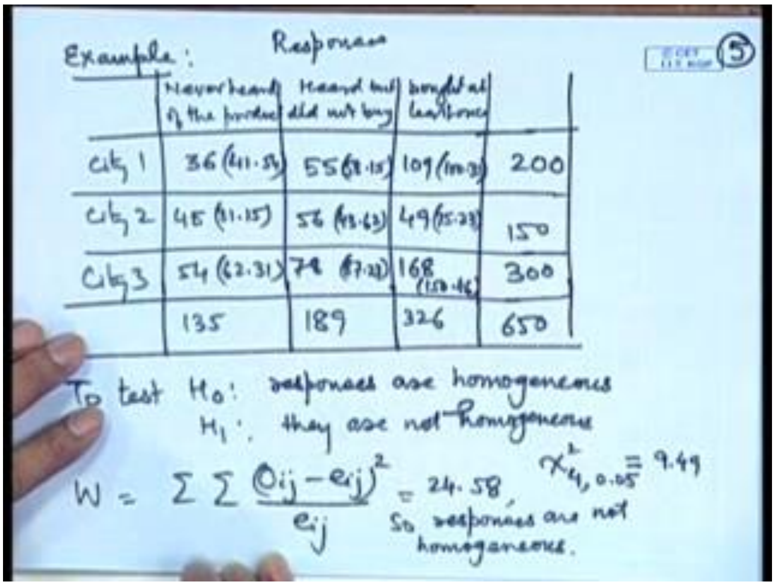
Test of homogeneity in $r \times c$ contingency tables:

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

So, in place of independence this is termed as test of homogeneity in r by c contingency tables. So, in this case we firstly stratify the population in say r or c categories and then we take samples of fixed sizes from each of the categories and see their classification with respect to other categories.

We want to test whether responses are homogeneous. Now you see the sampling condition has been slightly modified, in the earlier case we took full sample size and that means, for the full population we take a sample and then we see that to which i j th cell they fall. Now here either the row sums or the column sums are fixed, and then we see that what is the frequency of each cell in each row or each column; so the situation is slightly different, but the test of chi square goodness of fit which we have given for independence, the same test is valid here also let me give this through an example.

(Refer Slide Time: 13:56)



Example: Responses

	Never heard of the product	Heard but did not buy	bought at least once	
city 1	36 (41.54)	55 (58.15)	109 (100.31)	200
city 2	45 (31.15)	56 (43.62)	49 (75.23)	150
City 3	54 (62.31)	78 (87.23)	168 (150.46)	300
	135	189	326	650

To test H_0: responses are homogeneous

H_1: they are not homogeneous

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

So responses are not homogeneous.

So, a new product is introduced in the market and now we want to see whether it has the same level of effect in different towns of the country or different regions of the state or different areas of the city.

So, in this particular case we consider the response to the product in 3 different cities; that means, the customer or you can say the responder must have already purchased the product or you might have heard about the product, but not purchased or he might not have heard the product. So, the responses are based on a survey. Now here what we do we fixed a number of respondents in each city, rather then we merge the data of the 3 cities and then taking a random sample, from each city we take a fix sample size. So, let me present the data in the following form that. So, we have city 1, city 2 and city 3. We fix that we are taking 200 respondents from city 1, we are taking 150 respondents from city 2, and we are taking 300 respondents from city 3.

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

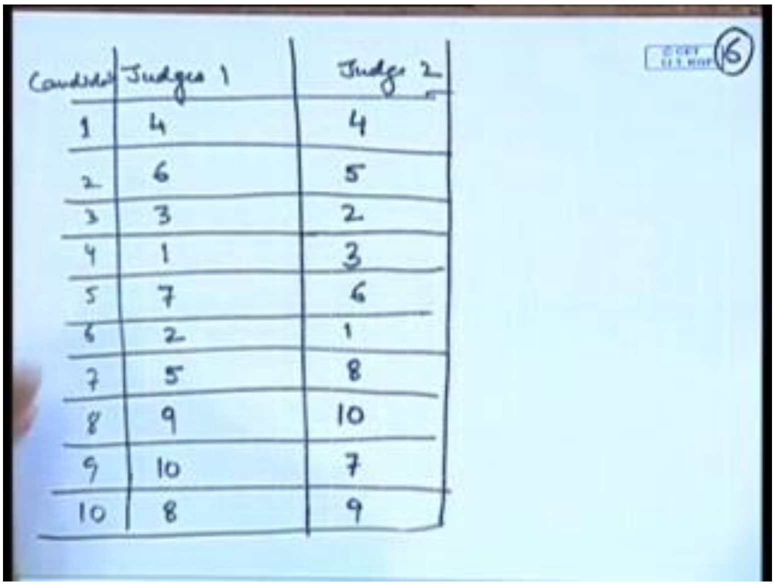
Now, to test whether the responses are homogeneous against the hypothesis they are not homogeneous. We apply the same test chi square test that is double summation o i j minus e i j square by e i j. So, we calculate the column sums this is 135, 189, 326 the total sample size is 650. So, based on this we can calculate like 200 into 135 divided by 165 that is say 41.54. 200 into 189 divided by 650 that is say 58.15 and so on 100.31 31.15, 43.62, 75.23, 62.31, 87.23, 150.46.

So, based on these calculations of o i j's and e i j's, you can evaluate this W and it terms out to be 24.58. Now if you look at chi square value here the degrees of freedom will be 3 minus 1 into 3 minus 1 that is 4. Suppose we take at 5 percent level of significance this value is 9.49. Naturally you can see the W is much larger than this. So, we conclude that H naught is rejected that means, responses are not homogeneous. So, you can see here that this chi square test for goodness of fit is applicable in various situations, we are able to test for goodness of fit; that means, whether the given data fits a given distribution we can test for independence of the 2 types of classifications in a contingency table, we may test for the pre homogeneity of the responses in a continuously table situation.

So, there are various applications let me we will come back to this again. Firstly, let me introduce 2 further measures of correlation earlier; we have seen the Karl Pearson measure of correlation, which is calculated as co variance divided by the standard deviations of the 2 variables. Now this measure of Karl Pearson it is completely dependent upon the numerical observations of the variables concerned for example, we may be looking at the relationship between the heights of say parents with the heights of the children, we may be concerned with the expenditure on the health care by families corresponding to their per capital income etcetera.

So, here actual measurements are required, but there are many situations in real life where the numerical values of the data are not very important. We may be simply concerned with say ranks of the values, or we may be concerned about the increasing or decreasing trend of the values. For example, 2 judges give ranks to a set of participants in a certain competition. So, now, they are not telling that for example, it could be selection procedure.

(Refer Slide Time: 21:33)



16

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

So, in a selection procedure suppose 10 candidates are there and there are 2 judges. So, judge 1 and judge 2. Now we have say candidates here say 1, 2, 3, 4, 5, 6, 7, 8, 9, 10.

So, rather than giving the scores their ranks are mentioned for example, judge 1 he ranks the candidate number say 4 as 1 whereas, judge 2 ranks say candidate number say 6 as 1; likewise they give ranks to all the candidates the candidate number 6 may be ranked 2 here. The candidate number 3 may be ranked 3, the candidate number 3 may be ranked 2 by this and this may be ranked 3, the candidate number 1 may be ranked 4, by say both of them the candidate number 2 may be ranked 6th here, and may be 5th here, candidate number 7th may be 5th here, 6th here this may be 7th here, say this may be 7th here may be this is 8th this could be 8th here, this could be 9th here may be this is 9, this is 10 this is 10.