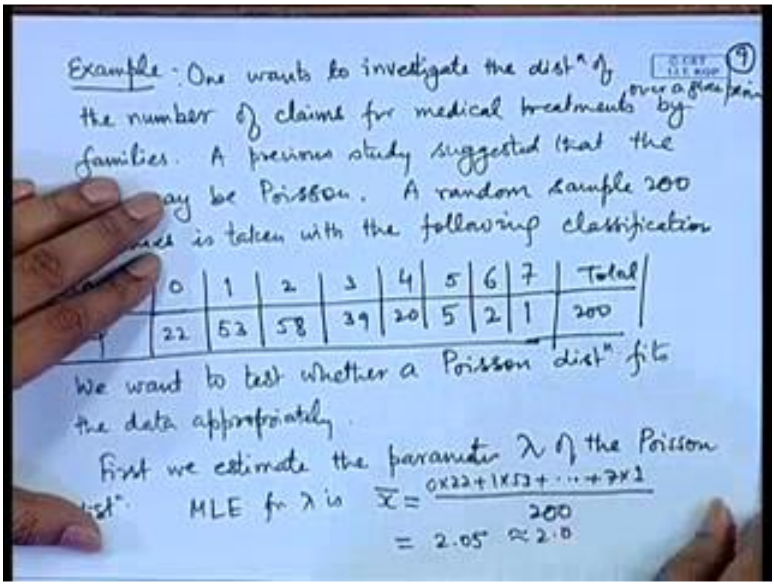
Probability and Statistics
Prof. Somesh Kumar
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture – 77
Chi-Square Test for Goodness Fit – II

(Refer Slide Time: 00:20)



9

Example: One wants to investigate the dist^n of the number of claims for medical treatments over a given period by families. A previous study suggested that the ay be Poisson. A random sample 200 is taken with the following classification

	0	1	2	3	4	5	6	7	Total
	22	53	58	39	20	5	2	1	200

We want to test whether a Poisson dist^n fits the data appropriately.

First we estimate the parameter λ of the Poisson dist^n. MLE for λ is $\bar{x} = \frac{0\times22+1\times53+\cdots+7\times1}{200}$

$= 2.05 \approx 2.0$

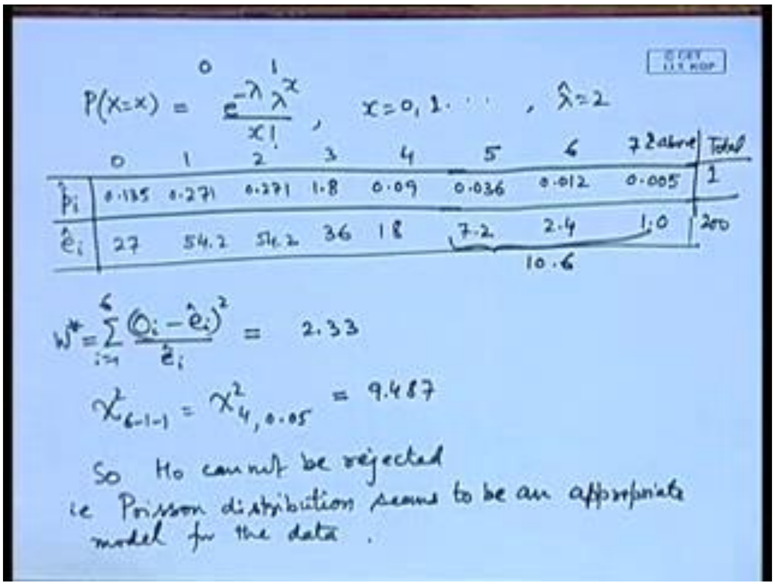
Let us take one case where the distribution will depend upon certain unknown parameter. So, one wants to investigate the distribution of the number of claims for medical treatments by families. A previous study suggested that the distribution may be Poisson. So to investigate this, a random sample of 200 families is taken with the following classification. So, that is for each family how many claims are there and that is what is the frequency; how many families made how many claims.

So, it turned out that 22 families did not make any claim, 53 families made 1 claim, 58 families made 2 claims, 39 families made 3 claims, 20 families made 4 claims, 5 families made 5 claims, 2 families made 6 claims, and 1 family made 7 claims. That means no family made more than 7 claims over a given period. Maybe say 5 years time or may be over a (Refer Time: 02:54) etcetera.

So, here we want to test whether; we want to test whether a Poisson distribution fits the data appropriately. Now here we do not even know the parameter lambda of the Poisson distribution, so firstly we will estimate that. So, first we estimate the parameter lambda of the Poisson distribution. Now we have already done estimation, we may use say maximum likelihood estimator or say minimum various unbiased estimator for lambda. So, for example, maximum likelihood estimator for lambda is x bar. So, x bar is the mean which can be evaluated from here that is 0 into 22 plus 1 into 53 and so on plus 7 into one divided by 200. The total number of families is 200. So, this value turns out to be 2.05.

So, we may approximately take two as the lambda value. And we like to check whether this data follows Poisson at the rate 2. Now this is a reasonable approximation, because 2.05 is the value and here we are talking about the number of claims. So, it is appropriate to take lambda to be an integral approximation of the value and it is extremely close, so this is fine.

(Refer Slide Time: 05:12)



$P(X=x) = \frac{\overset{0}{e}^{-\lambda}\,\overset{1}{\lambda}^{x}}{x!}$, $x = 0, 1 \cdots$, $\hat{\lambda} = 2$

	0	1	2	3	4	5	6	7 ≥ above	Total
p_i	0.135	0.271	0.271	1.8	0.09	0.036	0.012	0.005	1
$\hat{e}_i$	27	54.2	54.2	36	18	7.2	2.4	1.0	200
							10.6		

$$W^* = \sum_{i=1}^{6} \frac{(O_i - \hat{e}_i)^2}{\hat{e}_i} = 2.33$$

$$\chi^2_{6-1-1} = \chi^2_{4,0.05} = 9.487$$

So Ho cannot be rejected
ie Poisson distribution seems to be an appropriate model for the data .

So, we want to find out the probabilities of x is equal to say small x for 0, 1 etcetera. Now the formula in the passion distribution is e to the power minus lambda lambda to the power x by x factorial, x is equal to 0, 1, 2 and so on. Now the point here is that this is an infinite value distribution, here we will calculate the probabilities only for 0, 1 to 7; so they will not add up to 1. So, what we will do? We can calculate the probabilities 0, 1 to up to 6 and 7 we will put into the form of 7 and above. Although the data is not collected like that it was observed that for 8th there is no family which made 8 claims, there is no family which made 9 claims etcetera. So, basically the value 7 corresponds to 7 and above in that case the probability of the events each of the classes will be adding up to 1.

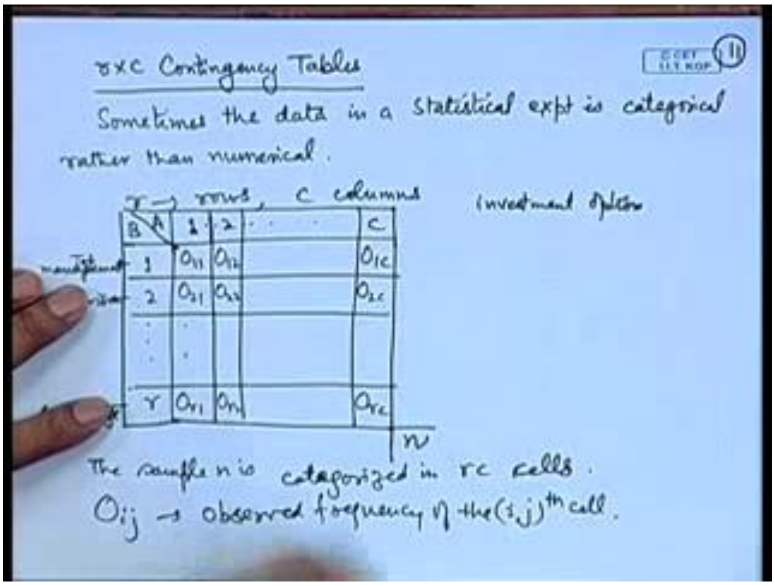
So, we calculate the probabilities and we report it like this. So, p i is for 0, 1, 2, 3, 4, 5, 6, 7 and above. So, substituting the value of lambda is lambda hat is equal to 2 we can calculate this is e to the power minus 2 that is equal to 0.135. The probability of x is equal to 1 that will become lambda e to the power minus lambda. So, that is twice e to the power minus lambda, so it is 0.27. This value is 0.271 again because this is again 2 to the power minus 2. This is 1.8, 0.09 then 0.036, 0.012, 0.005; the total is 1 here.

So, now the expected frequencies the estimate of the expected frequencies can be calculated by multiplying by 200 here. So, will get it as 20, 54.2, 36, 18, 7.2, 2.4 and 1.0; the total is 200. Now you note here the self frequency of 6 and 7 and above these are much below 5. In fact, even if you merge these two cells the expected frequency will be only 3.4. So, then this will not allow us to use the chi square approximation here. So, what we do we merge the last three cells. So, if you merge these last three cells and add up will get 10.6 as the expected frequency which is above 5.

So, in place of 8 cells now we have only 6 cells. So, sigma oi minus ei; so we will add to then merge the observed frequency also here, the observed frequency must will give 5 plus 2 plus 1 that is equal to 8. So, now, oi minus ei is square over ei 1 to 6 only, this is w is star. That is equal to 2.33. Now chi square value has to be on k minus m minus 1. So, here m is 1 because only 1 parameter is there; k is 6 so this becomes 6 minus 1 minus 1 that is chi square 4. Now if you see the value at say 0.05 that is 9.48 etcetera. So, H naught cannot be rejected; that is Poisson distribution seems to be an appropriate fit or appropriate model for the data, at least on the basis of the given data we have no reason to reject H naught that the distribution comes from a Poisson.

The chi square test for goodness of fit has other applications also. Let us consider the data which is not quantitative, but rather qualitative in nature.

(Refer Slide Time: 11:00)



r×c Contingency Tables

Sometimes the data in a statistical expt is categorical rather than numerical.

r → rows, c columns

investment options

B \ A	1	2	· · ·	c
1	O_{11}	O_{12}		O_{1c}
2	O_{21}	O_{22}		O_{2c}
⋮	⋮			
r	O_{r1}	O_{r2}		O_{rc}

n

The sample n is categorized in rc cells.

O_{ij} → observed frequency of the $(i,j)^{th}$ cell.

Let us consider the situation of contingency tables; r by c Contingency Tables. Sometimes the data in a statistical experiment is categorical rather than numerical. For example, in a population of individuals we will like to know how many of the people are smokers and how many of them are non smokers. So, it is a categorical data; that is there are two categories of person: one who are a smokers one who are non smokers.

We may also categorize them according to those who ultimately get lung cancer and who do not get the lung cancer. So now the situation is like this in a population we have characterized according to two different methods of categorization: one is the smoking habit and another is the incidence of a disease. Now we want to check whether there is any association between these two characteristics or two attributes. That means, is it true or is it found on the basis of the data that those who smoke they get they are more likely to get a lung cancer.

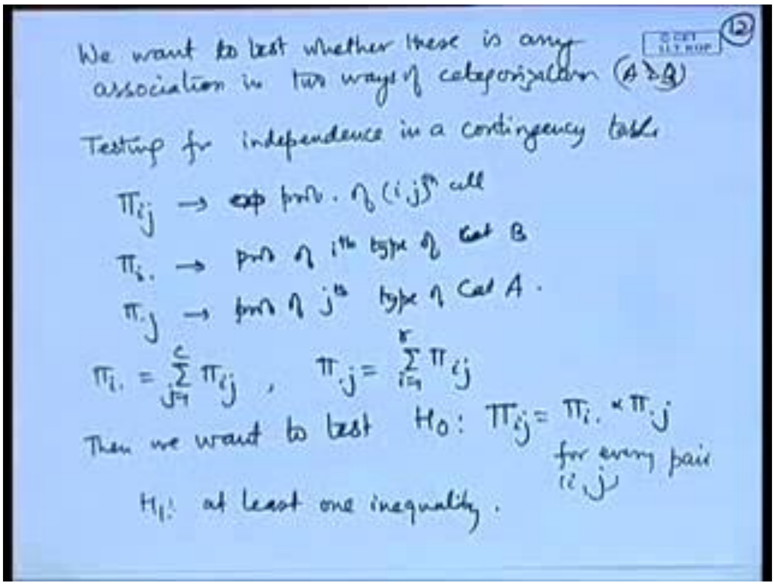
So, this is called testing for independence in a contingency table. So, what is the contingency table? That we have two types of categorizations: one is say we will represent in the form of rows and another type in the columns. So, we say r rows and c columns. So, the data may be represented like this. We have category A, we have category B. This may be 1, 2 up to c and here you will have 1, 2 up to r.

For example, in an organization people are classified according to their professional hierarchical levels. For example, one may be the top person of the organization that is say top management. In two we may have some people who are in the supervisory position or executives and here you may have say daily wage workers. And now we want to look at the attitude towards a certain option. So, for example, there is an option given for them for investment; investment option. That means, from their salary they may be allowed to invest in certain say shares of the company. And there may be say 4 different types of shares: one is where you get a fix kind of return, another is where you have a safe return; that means the equity is distributed into safe instruments, next may be some balanced; that is partially risk and partially safe, and there may be risky option.

So, may be you will find that out of a sample of n employees of the organization you found O 11 are of that is those who are in the top management and those who go for option one O 12, O 1c; O 21, O 22, O 2c; O r1, O r2, O rc. So this is the data. That is the sample n is categorized in rc cells and O ij is the observed frequency of the i, j-th cell.

What we want to test here is.

(Refer Slide Time: 16:33)



We want to test whether these is any association in two ways of categorization (A & B)

Testing for independence in a contingency table

$\pi_{ij} \rightarrow$ exp prob. of $(i,j)^{th}$ cell

$\pi_{i.} \rightarrow$ prob of i^{th} type of cat B

$\pi_{.j} \rightarrow$ prob of j^{th} type of cat A.

$\pi_{i.} = \sum_{j=1}^{c} \pi_{ij}$, $\pi_{.j} = \sum_{i=1}^{r} \pi_{ij}$

Then we want to test $H_0: \pi_{ij} = \pi_{i.} \times \pi_{.j}$ for every pair (i,j)

H_1: at least one inequality.

.

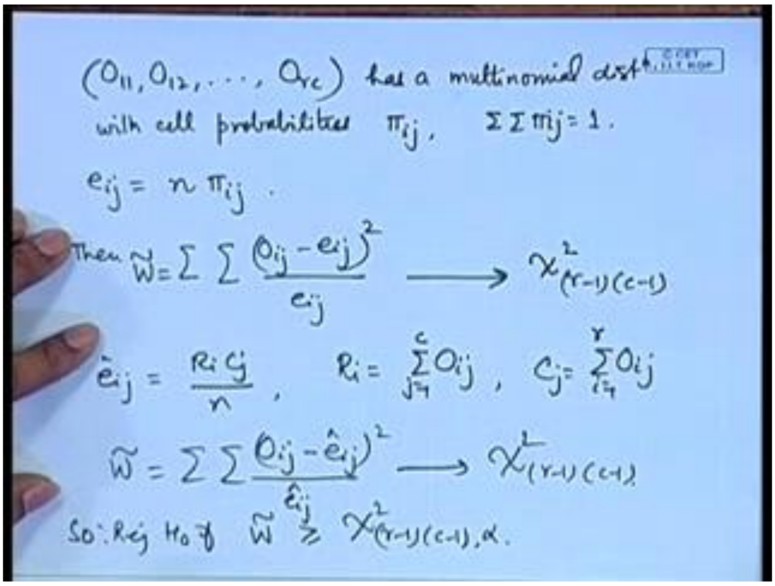
We want to test whether there is any association in two ways of categorization; that is A and B, whether A and B are having any association. As I just gave the example of smoking and caner, similarly here may be that those who are in the higher order in the hierarchy of the company they may like to invest in high risk equities etcetera, whereas the those who are in the lower income group or lower in the hierarchy they may like to go for safe this one. Suppose this could be our hypothesis or we may make a hypothesis that there is no relation.

So, this is called testing for independence in a contingency table. So, you can say here pi ij is the expected or you can say probability of i, j-th cell that is the theoretical probability. Pi i dot is the probability of i-th type of category say B. Pi dot j is the probability of j-th type of category A. So, actually here pi i dot will be equal to sigma pi ij- j is equal to 1 to c and pi dot j is actually sigma pi ij i is equal to 1 to r. Then we want to test H naught pi ij is equal to pi i dot into pi dot j; for every pair ij. And H 1 is at least one inequality.

Now if you see carefully this is nothing but a generalization of the test of goodness of fit itself. In the test of goodness of fit want we wanted to test is that whether the data comes from a particular distribution? The procedure adopted was that we divided the range of the distribution into k intervals or you can say k cells, and we looked at the observed frequencies; the theoretical distribution of the observed frequencies was multinomial.

Now likewise, here if you see what we are claiming here is that the probabilities are pi ij's for i, j-th cell, the observed frequency is O ij. So, if you look at the distribution of the categories here that is O 11, O 12, O 1c and so on up to o r c the joint distribution of this will again be multinomial. That means, the test will be actually of the previous form itself with little bit modification.

(Refer Slide Time: 20:43)



$(O_{11}, O_{12}, \ldots, O_{rc})$ has a multinomial distⁿ

with cell probabilities π_{ij}. $\sum\sum \pi_{ij} = 1$.

$e_{ij} = n\,\pi_{ij}$.

Then $\tilde{W} = \sum\sum \frac{(O_{ij} - e_{ij})^2}{e_{ij}} \longrightarrow \chi^2_{(r-1)(c-1)}$

$\hat{e}_{ij} = \frac{R_i C_j}{n}$, $R_i = \sum_{j=1}^{c} O_{ij}$, $C_j = \sum_{i=1}^{r} O_{ij}$

$\tilde{W} = \sum\sum \frac{(O_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} \longrightarrow \chi^2_{(r-1)(c-1)}$

So: Rej H_0 if $\tilde{W} \geq \chi^2_{(r-1)(c-1),\alpha}$.

So, if we write down O 11, O 12 and so on up to O rc, this has a multinomial distribution with cell probabilities pi ij with sigma of pi ij is equal to 1. So, using the previous argument if I define say e ij is equal to n times pi ij, then double summation O ij minus e ij square by e ij that is call it say w tilde; then this is asymptotically chi square distribution on r minus 1 into c minus 1 degrees of freedom. The values of e ij's are estimated by considering Ri Cj by n; where Ri is actually summation of O ij j is equal to 1 to c, and Cj is sigma O ij i is equal to 1 to r. So, if we make use of this O ij minus e ij hat square by e ij hat then this will be asymptotically chi square on r minus 1 c minus 1.

So, the test is rejects H naught if w tilde is greater than or equal to chi square r minus 1 c minus 1 alpha. So, the chi square test for goodness of fit is applicable for testing for independence in a r by c contingency table also. So, in the next class I will give some applications of this test we will also see that this is applicable to a slightly different situation. Here you can see that we are considering a random sample of size n from the population and then we are making this classification.

In certain other situations even this row totals or columns totals may be fixed and then the sampling is done form there. That means something like a stratified sampling. We will see that even in the stratified sampling the same formula is applicable. So, in the next class I will be covering those portions.