

Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology – Roorkee

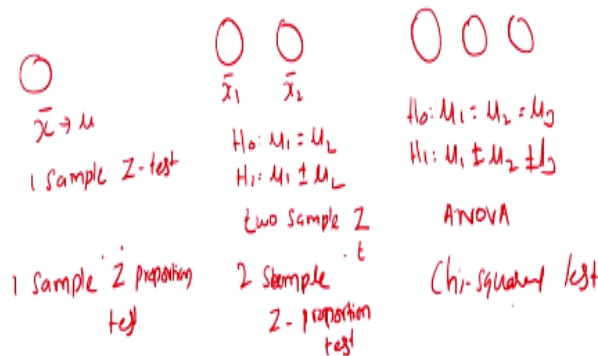
Lecture – 46
Chi-Square Test of Independence-I

Welcome students. Today we are going to see a new topic that is chi-square test. Chi-square test has 2 applications, one is to test the independence, second one is to test the goodness of fit. In this class, we are going to see the test of independence.

(Refer Slide Time: 00:44)

Agenda

- To understand χ^2 Test of Independence



The agenda for this class is to understand chi-square test of independence. Before going to the topic let us see when will you go for chi-square test. In the beginning of the lecture I have explained different types of data nominal, ordinal, interval ratio. Whenever the data is nominal or ordinal you have to go for your test called chi-square test because data which are nominal in nature you cannot go for Z test, you cannot go for T test or ANOVA, (()) (01:14) regression it cannot be done.

So now I will explain how this connection has with other test. For example, we might have seen we have studied one sample T test suppose there was a sample 1 was there we have seen with the help of \bar{x} we have predicted μ . After that we have seen 2 sample T test this is 1 sample Z test. Whenever there is a 2 population, population 1, population 2. See this is \bar{x}_1 this is \bar{x}_2 .

Here what we have compared what was our null hypothesis $\mu_1 = \mu_2$ $H_1 \mu_1 \neq \mu_2$. Here we are comparing 2 population at a time. So here what we have done we have done 2 sample we write we have done 2 samples Z test similarly we have done 2 sample T test also 1 this is 2. Suppose there are 3 population, population 1 and population 2 and population 3. Here suppose we want to compare the mean of this population.

Our null hypothesis will be $\mu_1 = \mu_2 = \mu_3$; alternative hypothesis is $\mu_1 \neq \mu_2$ $\neq \mu_3$. Here what we have done whenever we want to compare more than 2 population we have gone for ANOVA right. In the same way you see there is 1 sample here we can say 1 sample proportion test, 1 sample Z proportion test. Suppose if you want to compare the proportion of 2 population we can do 2 sample Z proportion test.

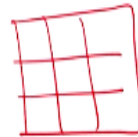
There may be a situation where you have to compare the proportion of more than 2 sample that time we should go for this chi-square test. Suppose if you are comparing mean where the population is more than 2 you should go for ANOVA. Suppose if you are comparing the proportion where you have to compare more than proportion of more than 2 population you should go for chi-square test that is the logic of using chi-square test.

Even if there are 2 sample mean instead of using 2 sample Z test you can use ANOVA also you will get the same result. Similarly, there are 2 sample proportions if you want to compare their proportion instead of using 2 sample proportion test you can do the chi-square test also you will get the same answer because this ANOVA and chi-square test is a generalized format. If you want to compare only 2 that time also you can use ANOVA okay.

(Refer Slide Time: 04:31)

χ^2 Test of Independence

- It is used to analyze the frequencies of two variables with multiple categories to determine whether the two variables are independent.
- Qualitative Variables
- Nominal Data



Now we will go to our topic today's topic that is the Test of Independence. This test is used to analyze the frequencies of two variables with multiple categories to determine whether the two variables are independent or not. So whenever there is a qualitative variables whenever the data is nominal data we should go for test of independence. Here, we are going to have 2 category suppose this is called table contingency table. There will be some value in row, some value in the column. So we are going to see whether these are dependent or independent with the help of an example the next I will explain.

(Refer Slide Time: 05:19)

χ^2 Test of Independence: Investment Example

- In which region of the country do you reside?
A. Northeast B. Midwest C. South D. West
- Which type of financial investment are you most likely to make today?
E. Stocks F. Bonds G. Treasury bills

		Type of financial Investment			
		E	F	G	
Geographic Region	A			O_{13}	n_A
	B				n_B
	C				n_C
	D				n_D
		n_E	n_F	n_G	N

For example, suppose you have conducted a questionnaire in that questionnaire this is an example of investment example you asked in which region of the country do you reside there was a 4 options. First option is Northeast, Midwest, South, West and you have asked another

question also which type of financial investment are you most likely to make today. The options were stocks, bonds and treasury bills.

This dataset I have captured in the form of a here table so this table is called contingency tables. See in rows I have captured the geographic regions say Northeast, Midwest, South, West. In column I have captured what type of financial investment they are going to make so that I have E, F, G. Suppose I wanted to make an assumption I wanted to test is there any connection between, is there any dependency between the geographic regions where they reside and the type of investment they are willing to make.

So there are 2 variables; one is geographic region another variable is type of financial investment whether these are dependent or independent. So this kind of examples, this kind of problems can be solved with the help of this chi-square test okay. So what is your null hypothesis the geographic regions and the type of investment which they makes are independent there is no connection. Alternative hypothesis it is not independent there is a dependency.

(Refer Slide Time: 07:11)

χ^2 Test of Independence: Investment Example

If A and F are independent,
 $P(A \cap F) = P(A) \cdot P(F)$

$$P(A) = \frac{n_A}{N} \quad P(F) = \frac{n_F}{N}$$

$$P(A \cap F) = \frac{n_A}{N} \cdot \frac{n_F}{N}$$

$$e_{AF} = N \cdot P(A \cap F)$$

$$= \cancel{N} \left(\frac{n_A}{\cancel{N}} \cdot \frac{n_F}{N} \right)$$

$$= \frac{n_A \cdot n_F}{N}$$

		Type of Financial Investment			
		E	F	G	
Contingency Table Geographic Region	A		e_{12}		n_A
	B				n_B
	C				n_C
	D				n_D
		n_E	n_F	n_G	N

This was a theory behind this test of independence. Suppose if A and F are independent A, F. What is the A here A means the first option that was he belongs to Northeast region, F means he is willing to invest in the bond. If A and B are independent event, we can write P of A intersection F is P of A and P of F then we will find out what is P of A. P of A is your n of A / total N capital N that is number of sum of all the elements.

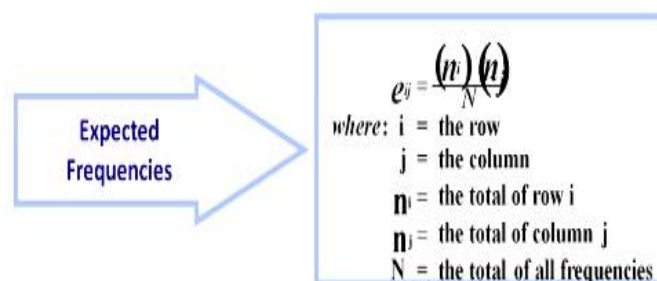
Then P of F is n of F divided by total element. So when you multiply intersection of P of A and F so that will be n of A divided by n multiplied by n of F divided by n. Now you will get here in the terms of probability, but if you want to get in terms of frequencies so that has to be multiplied by your capital N. So expected value of AF = N multiplied by P A intersection F.

So it is N. For P of A we know it is n of A / n P of F we know it is n of F / N. So this N N gets cancelled the remaining is n of A dot n of F / N. So this is very frequent formula which we are going to use. Suppose if you wanted to know this expected values in the cells what you have to do. You have to multiply the row sum and column sum divided by your capital N.

So row sum multiplied by column sum divided by total number elements that will give you the expected value of each cell. One more thing you see that we are multiplying by N here N is because if you put P of A intersection F we will get only probability if you want to get the answer in terms of frequency that has to be multiplied by your N that is why we are getting this N.

(Refer Slide Time: 09:21)

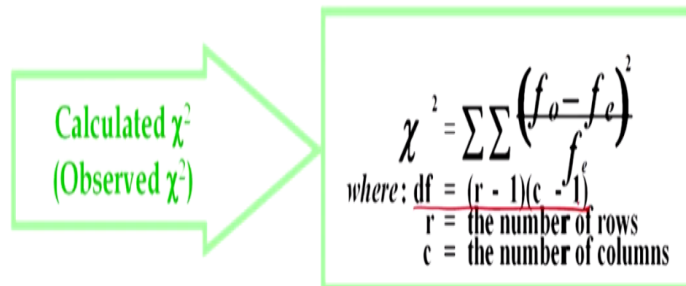
χ^2 Test of Independence: Formulas



So expected frequency is e_{ij} is $n_i n_j$ i represents row j represents the column n_i means the total of row i multiplied by n_j the total of column j divided by capital N the total of all frequencies. This is our expected frequencies. Obviously there is one more frequency which you have to find out the observed frequency that will be given in the problem itself.

(Refer Slide Time: 09:52)

χ^2 Test of Independence: Formulas


$$\chi^2 = \sum \sum \frac{(f_o - f_e)^2}{f_e}$$

where: $df = (r - 1)(c - 1)$
 r = the number of rows
 c = the number of columns

Then how to find out the test statistics of our chi-square so the calculated chi square our observed chi-square value is the sigma of sigma see first one is observed frequency f_o minus expected frequency whole square / expected frequency. You see that this square is not for this expected frequency that is only for the numerator and another important thing is degrees of freedom.

The degrees of freedom is row - 1 multiplied by column - 1 that means that many number of independent cells we can supply any values. So r represents number of rows c represents number of columns.

(Refer Slide Time: 10:40)

Example for Independence

We will take one example using the theory which I have taught you so far. We will solve a problem of test of independence.

(Refer Slide Time: 10:49)

χ^2 Test of Independence

H_0 : Type of gasoline is independent of income
 H_a : Type of gasoline is not independent of income

Suppose before starting any hypothesis testing problem the first step is to first write the null hypothesis. The null hypothesis is the type of gasoline is independent of income. Alternative hypothesis is type of gasoline is not independent of income. Generally, there are different type of gasoline we have a assumption that people were having higher income they will go for good quality in fuel. So we are going to test this is there is any connection, is there is any dependency between their level of income and the type of gasoline they prefer.

(Refer Slide Time: 11:33)

χ^2 Test of Independence

$r = 4$	Type of Gasoline		
	$c = 3$	Regular	Premium
Income			
Less than \$30,000			
\$30,000 to \$49,999			
\$50,000 to \$99,000			
At least \$100,000			

This was our problem setup. So in the rows I have captured their level of income less than 30,000 dollar next category is \$30,000 to \$49,999 next is \$50,000 to \$99,000 it is more than \$100,000. In column I have asked what type of gasoline you are using, fuel you are using

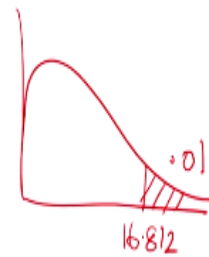
regular, premium, extra premium. Generally, there is an assumption whenever the income level is increases the people may go for good quality fuel.

So there is a dependency that is our assumption, there may be a dependency between their level of income and the type of fuel which they use. Here the rows there are 4 rows 1, 2, 3, 4 there are 3 columns so $r = 4$ $c = 3$.

(Refer Slide Time: 12:40)

χ^2 Test of Independence: Gasoline Preference Versus Income Category

$$\begin{aligned}\alpha &= .01 \\ df &= (r - 1)(c - 1) \\ &= (4 - 1)(3 - 1) \\ &= 6 \\ \chi^2_{.01,6} &= 16.812\end{aligned}$$



$$\begin{aligned}\text{If } \chi^2_{\text{cal}} > 16.812, & \text{ reject } H_0. \\ \text{If } \chi^2_{\text{cal}} \leq 16.812, & \text{ do not reject } H_0.\end{aligned}$$

First we will find out the degrees of freedom. Assume that the alpha is 1 percentage the degrees of freedom is row - 1 multiplied by column - 1 there are 4 rows 4 - 1 so 4 - 1 is 3 there are 3 column 3 - 1 = 2 so 3 into 2 it is 6. So for 6 degrees of freedom so the chi-square distribution is the $\chi^2(6)$ (13:05) distribution it will be like this. So when right side area is 1 percentage for 6 degrees of freedom the value which got from the table is 16.812. The next slide I will tell you this value we can find out with the help of python.

This was the value which we got from the table. Next what we have to do, we have to calculate the chi square value if our calculated chi-square value how we will calculate it using this formula observed frequency - expected frequency whole square divided by expected frequency. Using this formula, we have to find out the chi-square calculated. If that value is greater than 16.82 we will reject our null hypothesis. If it is less than 16.812 we will accept our null hypothesis.

(Refer Slide Time: 14:06)

Python code

```
In [5]: import pandas
import numpy
from scipy import stats
```

```
In [6]: stats.chi2.ppf(0.99,6)
```

```
Out[6]: 16.811893829770927
```

As I told you with the help of python import pandas, import numpy from scipy import stats stats.chi2.ppf when it is a 0.99 because if we want to know one percentage it is 0.99 for 6 degrees of freedom this is 16.811 that sort of value we got it.

(Refer Slide Time: 14:27)

Gasoline Preference Versus Income Category: Observed Frequencies

Income	Type of Gasoline			
	Regular	Premium	Extra Premium	
Less than \$30,000	85	16	6	107
\$30,000 to \$49,999	102	27	13	142
\$50,000 to \$99,000	36	22	15	73
At least \$100,000	15	23	25	63
	238	88	59	385

This was the data which is given is what is the value which is their inside the cell it is called observed frequency. So what is the meaning of this 85 those were having income less than \$30,000 they have gone for regular type of gasoline. What is the premium, what is the interpretation of this 16 those were having income less than \$30,000 only 16 people have gone for premium type right this one.

You see that when the level of income is increasing you see that here also the number is the people gone for extra premium also increasing. It seems to be (()) (15:06) dependency

between their level of income and type of gasoline they choose. So this is the given data which we have captured. So the first step is we have to find out the row total. The first row total is 107 second row total is 142 third row 73, fourth row 63. Then finding the column total the first column total is 238, second column total is 88, third column total is 59. The value which are given is called observed frequency.

(Refer Slide Time: 15:47)

Gasoline Preference Versus Income Category: Expected Frequencies

Income	Type of Gasoline			
	Regular	Premium	Extra Premium	
Less than \$30,000	(66.15) 85	(24.46) 16	(16.40) 6	107
\$30,000 to \$49,999	(87.78) 102	(32.46) 27	(21.76) 13	142
\$50,000 to \$99,000	(45.13) 36	(16.69) 22	(11.19) 15	73
At least \$100,000	(38.95) 15	(14.40) 23	(9.65) 25	63
	238	88	59	385

Now we should go for our expected frequency here the expected frequency value is given in the bracket. For example, how we got this 66.15 this 66.15 is nothing, but multiplication of row total 107 and the column total 238 divided by 385. So that value is nothing, but 66.15 that is given in the bracket. So the values which are given in the bracket it is called expected frequency.

The value which is not in the bracket it is called observed frequency that is the data which are given to us. So for the second dataset how we have got 24.46 so row total 107 column total 88 / 388 so we will get 24.46. For third one row total 107 column total 59 the total value is 385 like this we have to find out all the cells, all the cells after finding which was given in the bracket. Now we will go for chi-square calculated value.

(Refer Slide Time: 17:01)

Gasoline Preference Versus Income Category: χ^2 Calculation

$$\chi^2 = \sum \sum \left(\frac{f_o - f_e}{f_e} \right)^2$$

$$= \frac{(85 - 66.15)^2}{66.15} + \frac{(16 - 24.46)^2}{24.46} + \frac{(6 - 16.40)^2}{16.40} +$$

$$\frac{(102 - 87.78)^2}{87.78} + \frac{(27 - 32.46)^2}{32.46} + \frac{(13 - 21.76)^2}{21.76} +$$

$$\frac{(36 - 45.13)^2}{45.13} + \frac{(22 - 16.69)^2}{16.69} + \frac{(15 - 11.19)^2}{11.19} +$$

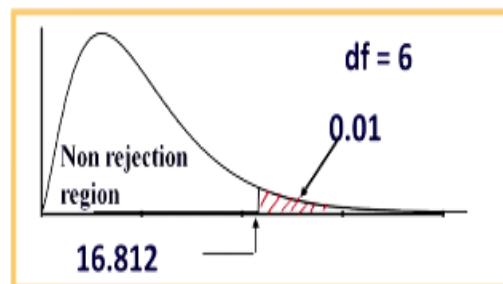
$$\frac{(15 - 38.95)^2}{38.95} + \frac{(23 - 14.40)^2}{14.40} + \frac{(25 - 9.65)^2}{9.65}$$

$$= 70.75$$

Now we will find out the chi-square calculated value as we know that the formula for chi-square calculated value is observed frequency – expected frequency divided by expected frequency. So (()) (17:12) 85 – 66.15 whole square / 66.5 it is the first cell first row first cell. First row second cells this 16 is your observed value this is 24.46 is our expected value whole square divided by 24.46. So if you keep on extent this you can go up to last cell so when you sum it, it is coming 70.75 this is 5 this is 70.75.

(Refer Slide Time: 17:44)

Gasoline Preference Versus Income Category: Conclusion



$$\chi_{Cal}^2 = 70.75 > 16.812, \text{ reject } H_0.$$

We know that previously we have seen that this value is 0.01 our calculated value is 70.75 it is lying on the right hand side so we are going to reject our null hypothesis. When you reject our null hypothesis what was null hypothesis their level of income and the type of fuel they choose are independent. So when you reject it what we are concluding there is a dependency between their level of income and the type of fuel they choose. Generally, it is an assumption

when the level of income is increasing they will go for higher quality of the fuel that was the conclusion.

(Refer Slide Time: 18:28)

Contingency Tables

Contingency Tables

- Useful in situations involving multiple population proportions
- Used to classify sample observations according to two or more characteristics
- Also called a cross-classification table.

The table which we have seen previously it is called a contingency table. It is useful in situations involving multiple population proportions. It is used to classify sample observations according to 2 or more characteristics also called cross-classification table another name for contingency table is cross-classification table.

(Refer Slide Time: 18:52)

Contingency Table Example

Hand Preference vs. Gender

Dominant Hand: Left vs. Right

Gender: Male vs. Female

- 2 categories for each variable, so the table is called a 2 x 2 table
- Suppose we examine a sample of 300 college students

We will solve one example here. It also is similar to our previous problem there is another example. Suppose we are going to compare the hand preference versus gender. So the dominant hand maybe for some people may be left some people is right. The gender is male versus female we are going to have here hypothesis that is there any connection, is there any

dependency between the gender and their dominant hand. So we have 2 categories for each variable. So this is called 2 cross 2 table. We examine the sample of 300 college students this was the outcome.

(Refer Slide Time: 19:35)

Contingency Table Example

Sample results organized in a contingency table:

sample size = n = 300:
 120 Females, 12 were left handed
 180 Males, 24 were left handed

Hand Preference	Gender		
	Female	Male	
Left	12	24	36
Right	108	156	264
	120	180	300

In the rows we have asked are you left hand or right hand dominant hand say left in the column we have captured the gender female and male. There are 300 observations out of 300 observations 120 are female 180 are male. Out of 300 36 are left handed, 264 are right handed. So the sample result organized in a contingency table. The sample size is n so 120 females, 20 were left handed this one, 180 males, 24 were left handed right.

(Refer Slide Time: 20:20)

Contingency Table Example

$H_0: \pi_1 = \pi_2$ (Proportion of females who are left handed is equal to the proportion of males who are left handed)

$H_1: \pi_1 \neq \pi_2$ (The two proportions are not the same Hand preference is **not** independent of gender)

- If H_0 is true, then the proportion of left-handed females should be the same as the proportion of left-handed males.
- The two proportions above should be the same as the proportion of left-handed people overall.

So what is our hypothesis $H_0 = \pi_1 = \pi_2$. The proportion of females who are left handed is equal to the proportion of male who are left handed. Now dominant hand left hand is taken as

the reference we are going to compare that left hand people with respect to their gender. So taking left hand is a dominant hand we are going to find out is there is connection between their hand dominant and their gender that is a null hypothesis.

Null hypothesis proportion of female who are left handed is equal to the proportion of male who are left handed. Suppose if you accept null hypothesis so there is no connection between their dominance of left hand and their gender. What is alternate hypothesis? The two proportions are not the same hand preference is not independent of gender. So what will happen if H0 is true then the proportion of left handed female should be there.

Same as the proportion of left handed males. So we can say there is no dependency. The two proportions above should be the same as the proportion of left handed people overall instead of taking left hand as a reference you can take the right hand also both result will be the same. **(Refer Slide Time: 21:47)**

The Chi-Square Test Statistic

The Chi-square test statistic is:

$$\chi^2 = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e}$$

where:

f_o = observed frequency in a particular cell

f_e = expected frequency in a particular cell if H_0 is true

χ^2 for the 2 x 2 case has 1 degree of freedom

Assumed: each cell in the contingency table has expected frequency of at least 5

We have seen this formula that is a chi-square test statistics is observed frequency minus expected frequency whole square divided by expected frequency. f_o says observed frequency f_e is expected frequency. Here there are 2 cross 2 table so the degrees of freedom is $2 - 1$ multiple by $2 - 1$ so it is 1 degrees of freedom. We assume that there is an important assumption each cell in the contingency table has the expected frequencies at least 5. We have to make sure that the expected frequency is at least 5 that is an assumption if it is not there we have to collapse 2, 3 column so that to get the expected frequency is 5.

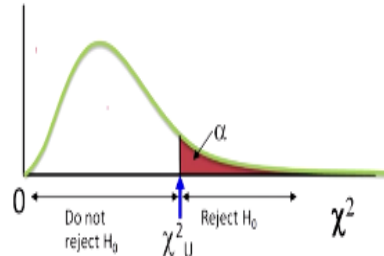
(Refer Slide Time: 22:36)

The Chi-Square Test Statistic

The χ^2 test statistic approximately follows a chi-square distribution with one degree of freedom

Decision Rule:

If $\chi^2 > \chi^2_U$, reject H_0 ,
otherwise, do not reject H_0



The chi-square test statistic approximately follows the chi-square distribution with one degrees of freedom what will happen the decision rule is if the chi-square value is greater than this limit we will reject it otherwise we will accept it otherwise (()) (22:51) do not reject it.

(Refer Slide Time: 22:54)

Observed vs. Expected Frequencies

Hand Preference	Gender		
	Female	Male	
Left	Observed = 12 ✓ Expected = 14.4 <small>$\frac{36 \times 120}{300}$</small>	Observed = 24 Expected = 21.6 <small>$\frac{36 \times 180}{300}$</small>	36
Right	Observed = 108 Expected = 105.6 <small>$\frac{264 \times 120}{300}$</small>	Observed = 156 Expected = 158.4 <small>$\frac{264 \times 180}{300}$</small>	264
	120	180	300

So this is the observed frequency. Now we have to find out for each cell expected frequency. How we got this expected frequency is nothing but 36 multiplied by 120 divided by 300. So how we got this one 36 multiplied by 124 divided by 300. How we got this value 36 multiple by 180 divided by 300 the same way how we got here 264 multiple by 120 divided not 120 300 here 264 multiplied by 180 divided by 300. So we will get here value for this.

(Refer Slide Time: 23:50)

The Chi-Square Test Statistic

Hand Preference	Gender		
	Female	Male	
Left	Observed = 12 Expected = 14.4	Observed = 24 ✓ Expected = 21.6	36
Right	Observed = 108 Expected = 105.6	Observed = 156 Expected = 158.4	264
	120	180	300

The test statistic is:

$$\chi^2 = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(12 - 14.4)^2}{14.4} + \frac{(108 - 105.6)^2}{105.6} + \frac{(24 - 21.6)^2}{21.6} + \frac{(156 - 158.4)^2}{158.4} = 0.7576$$

The next one we have to find out the observed frequency minus expected frequency whole square divided by expected frequency plus for this one $108 - 105.6$ whole square / 105.6 + for this cell it is $24 - 21.6 / 21.6$ whole square not whole square only for the numerator + $156 - 158.4$ whole square / 158.4 that is we are getting this 0.7576.

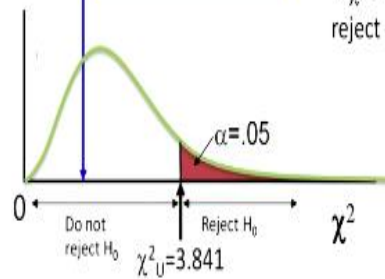
(Refer Slide Time: 24:28)

The Chi-Square Test Statistic

The test statistic is $\chi^2 = 0.7576$, χ^2_U with 1 d.f. = 3.841

Decision Rule:

If $\chi^2 > 3.841$, reject H_0 , otherwise, do not reject H_0



Here,
 $\chi^2 = 0.7576 < \chi^2_U = 3.841$,
 so you do not reject H_0 and
 conclude that there is
 insufficient evidence that the
 two proportions are different.

Now we have to mark this one so point the table value which we got this is chi-square calculated value the table values which for one degrees of freedom this is 3.814 our calculated value is lying on the acceptance side. So we have to accept null hypothesis. If chi-square value is greater than 3.841 (()) (24:51) otherwise do not reject it here the chi-square value that is 07576 is less than your 3.841.

You do not reject H_0 and conclude that there is insufficient evidence that the 2 proportions are different that means that both are same $P_1 = P_2$.

(Refer Slide Time: 25:17)

χ^2 Test for The Differences Among More Than Two Proportions

- Extend the χ^2 test to the case with more than two independent populations:

$$H_0: \pi_1 = \pi_2 = \dots = \pi_c$$

$$H_1: \text{Not all of the } \pi_j \text{ are equal } (j = 1, 2, \dots, c)$$

There we have compared only 2 proportions there may be a possibility we have to compare more than 2 proportions for example 3 proportions that case we will see in this problem. Extend the chi-square test to the case where with more than 2 independent populations say null hypothesis can be $\pi_1 = \pi_2 = \pi_3$ the alternate hypothesis not all of the proportions are equal.

(Refer Slide Time: 25:44)

The Chi-Square Test Statistic

The Chi-square test statistic is:

$$\chi^2 = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e}$$

where:

- f_o = observed frequency in a particular cell of the $2 \times c$ table
- f_e = expected frequency in a particular cell if H_0 is true
- χ^2 for the $2 \times c$ case has $(2-1)(c-1) = c - 1$ degrees of freedom

Assumed: each cell in the contingency table has expected frequency of at least 5

This other formulas are same as usual f_o is observed frequency f_e is expected frequency chi-square is the degrees of freedom. So the formula number of row – 1 multiplied by number of column – 1.

(Refer Slide Time: 25:58)

χ^2 Test with More Than Two Proportions: Example

The sharing of patient records is a controversial issue in health care. A survey of 500 respondents asked whether they objected to their records being shared by insurance companies, by pharmacies, and by medical researchers. The results are summarized on the following table:

We will see one example the sharing of patients records is a controversial issue in health care. A survey of 500 respondents asked whether they objected to their record being shared by insurance companies, pharmacies and by medical researchers. The results are summarized on the following table. So there are 3 category now one is whether they are objected to share the data for insurance companies, pharmacies and medical researcher.

(Refer Slide Time: 26:29)

χ^2 Test with More Than Two Proportions: Example

Object to Record Sharing	Organization			Row Sum
	Insurance Companies	Pharmacies	Medical Researchers	
Yes	410 <small>1040×500</small>	295 <small>1040×500</small>	335	1040 →
No	90 <small>1500</small>	205 <small>1500</small>	165	460 →
Column Sum	500 ↓	500 ↓	500 ↓	1500

So this table shows like this. So you see that 410 patients have objected to share their data with the insurance companies, 295 patients have objected to share their data with the pharmacies, 335 people have objected their data to share with the medical researchers. Now we have to find out whether the proportion of objection for sharing their data all these 3 categories are same or not. So we can say this is π_1 , this is π_2 , this is π_3 .

So null hypothesis will be $\pi_1 = \pi_2 = \pi_3$ that means that the people are always object to share their data irrespective of what kind of company it is that is our null hypothesis. There is a independency between sharing their data and the types of companies which they ask for the data. Here what you have done I have done the row sum this was the row sum this was second row sum then I found the column sum there are 3 columns. This data is our observed frequency 295, 410, 335.

(Refer Slide Time: 27:54)

χ^2 Test with More Than Two Proportions: Example

The overall proportion is:

$$p = \frac{X_1 + X_2 + \dots + X_c}{n_1 + n_2 + \dots + n_c} = \frac{410 + 295 + 335}{500 + 500 + 500} = 0.6933$$

Object to Record Sharing	Organization		
	Insurance Companies	Pharmacies	Medical Researchers
Yes	$f_o = 410$ $f_e = 346.667$	$f_o = 295$ $f_e = 346.667$	$f_o = 335$ $f_e = 346.667$
No	$f_o = 90$ $f_e = 153.333$	$f_o = 205$ $f_e = 153.333$	$f_o = 165$ $f_e = 153.333$

Next one from the observed frequency I have to find out the expected frequency. We have already observed frequency how will you find the expected frequency for example here 1040 multiplied by 500 divided by 1,500. So that value it will be 346.667. Similarly, for second dataset it is 1,040 multiplied by 500 divided by 1,500 that data is about 346. This way you can find out the expected frequency.

(Refer Slide Time: 28:40)

χ² Test with More Than Two Proportions: Example

Object to Record Sharing	Organization		
	Insurance Companies	Pharmacies	Medical Researchers
Yes	$\frac{(f_o - f_e)^2}{f_e} = 11.571$	$\frac{(f_o - f_e)^2}{f_e} = 7.700$	$\frac{(f_o - f_e)^2}{f_e} = 0.3926$
No	$\frac{(f_o - f_e)^2}{f_e} = 26.159$	$\frac{(f_o - f_e)^2}{f_e} = 17.409$	$\frac{(f_o - f_e)^2}{f_e} = 0.888$

The Chi-square test statistic is:

$$\chi^2 = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e} = 64.1196$$

Now I have given the final answer for the first cell that is the observed frequency minus expected frequency whole square divided by expected frequency for this cell it is 11.157, here it is 7.77 this is 0.392, this is 26.159, this is 17.409, this is 0.88. When you add it that value is your 64.1196.

(Refer Slide Time: 29:05)

χ² Test with More Than Two Proportions: Example

$$H_0: \pi_1 = \pi_2 = \pi_3 \quad \checkmark$$

$$H_1: \text{Not all of the } \pi_j \text{ are equal (j = 1, 2, 3)}$$

Decision Rule:

If $\chi^2 > \chi^2_{\alpha}$, reject H_0 , otherwise, do not reject H_0

$\chi^2_{\alpha} = 5.991$ is from the chi-square distribution with 2 degrees of freedom.

$$(2-1)(3-1) = 1 \times 2 = 2$$

Conclusion: Since $64.1196 > 5.991$, you reject H_0 and you conclude that at least one proportion of respondents who object to their records being shared is different across the three organizations

So what is a null hypothesis as I told you $\pi_1 = \pi_2 = \pi_3$ alternate hypothesis all of the π_j are equal that is j 1, 2, 3, 4. Decision rules if the calculated chi-square value is greater than your table value reject H_0 otherwise do not reject it. The table value which we got from the table is 5.9 what is the degrees of freedom for knowing this. You see that there are 2 rows is there so $2 - 1$ there are 3 column is there so $3 - 1$.

So this is 1 multiplied by 2 it will be 2 degrees of freedom. For 2 degrees of freedom for given alpha value the chi-square value which you got from the table is 5.991, but you see that our calculated value is our 64.116. So it is bigger than our table value so we have to reject and we can conclude that at least one proportion of the responds to object to their record being shared it is different across the 3 organizations.

So what will happen when you reject to a null hypothesis we can say it is not always equal there are somewhere it is not equal. So not all of the p_{ij} are equal. In this lecture, we started a new topic that is a chi-square test. Chi-square test has 2 applications. One is test of independence and goodness of fit. Today we have started with a test of independence I have taken a small example.

I have explained with the help of example how to test the test of independence. In the next class we will take one small problem with the help of python I will explain how to construct the contingency table. After constructing contingency table how to do chi-square test of independence using python that we will see in the next class. Thank you.