

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

Lecture – 47
Chi-Square Test of Independence-II

In the last class we started about the chi-square distribution. As I told you chi-square distribution having 2 application one is to test the independence, second one is the goodness of fit. We have seen in the previous class one example of how to test independence of the 2 variables. In this class, we will continue with that we will take another example we will solve with the help of python then we will go for another application of chi-square test that is the goodness of fit.

(Refer Slide Time: 00:56)

Agenda

- Using python to test the independence of variables
- Understanding goodness of fit test for Poisson

So agenda for this lecture is using python to test the independence of the variables. We will start another new topic that is application of chi-square distribution on testing the goodness of fit. First we will test the Poisson distribution.

(Refer Slide Time: 01:13)

Example

- Record of 50 students studying in ABN School is taken at random, the first 10 entries are like this:

res_num	aa	pe	sm	ae	r	g	c
1	99	19	1	2	0	0	1
2	46	12	0	0	0	0	0
3	57	15	1	1	0	0	0
4	94	18	2	2	1	1	1
5	82	13	2	1	1	1	1
6	59	12	0	0	2	0	0
7	61	12	1	2	0	0	0
8	29	9	0	0	1	1	0
9	36	13	1	1	0	0	0
10	91	16	2	2	1	1	0

This is an example record of 50 students studying say ABN school is taken at random. The first 10 entries are shown in like this. Why I have taken this example is in the previous lecture we have got the contingency table directly. The contingency table is given to you once the contingency table is given finding observed frequency and expected frequency then finding chi-square calculated values are very simple.

But in practical the chi-square distribution will not be given to you directly only excel data file or some file from database will be given. You have to form the contingency table after forming the contingency table then you should go for chi-square test. So in this example I have taken one hypothetical problem and there are 1, 2, 3, 4, 5, 6, 7 variables are there. The first variable is academic ability aa.

Second variable is parent education, the third variable is student motivation, the fourth variable is advisory evaluation, the next variable is religion, next variable is gender the last variable is community type. This dataset is to identify what are the variable that affect academic performance of a candidate. So here the academic ability is nothing, but test conducted for out of 100 marks.

So this is marks obtained by the candidates only I have shown only 10 dataset like that there are 50 dataset is there where I am show you when I am showing the python demo. The first one is the marks obtained by a student that is called academic ability so that is higher the marks higher the academic ability. The next variable is parent education the question is asked to the parent how many years you spend on the schooling.

Generally, parent education is a categorical variable, but instead of capturing categorical variable we have got that variable in the form of interval kind of a continuous variable. What we have asked to the parent that how many years you have spend in the schooling say 5 years, 6 years and so on. So now the parent education will become a continuous variable then student motivation.

We have asked the student suppose if you want to study if you want to get more marks are you willing to spend extra time for the study 1 means no, 0 means not decided, 2 means yes advisory evaluation. Advisory is like kind of a faculty advisor that faculty advisor can say whether this fellow will pass in the examination or he will do good performance in the examination.

1 means we will not do 0 means not decided, 2 means we will do good performance in the examination then r is the religion. There are 3 category of the religion 0, 1, 2 this is the explanation of the variables.

(Refer Slide Time: 04:16)

Example

Here :

- res_num = registration no.
- aa= academic ability
- pe = parent education
- sm = student motivation
- r = religion
- g = gender

For example, as I told you the first column is res respondent underscore number kind of a registration number aa is academic ability, pe is parent education, sm is student motivation, r is the religion, g is the gender.

(Refer Slide Time: 04:35)

Python code

```
In [1]: import pandas as pd
import numpy as np

In [2]: acad = pd.read_csv('AcademicAbilityData.csv')

In [3]: acad
Out[3]:
```

	rel_eva	as	pe	sm	ae	r	g	r
0	1	89	10	1	2	0	0	1
1	2	45	12	0	3	0	0	0
2	1	57	15	1	1	0	0	0
3	4	94	10	2	2	1	1	1
4	5	82	13	2	1	1	1	1
5	8	59	12	0	3	2	0	0
6	7	81	12	1	2	0	0	0
7	3	29	9	0	3	1	1	0
8	8	30	13	1	1	0	0	0
9	10	91	15	2	2	1	1	0
10	11	55	10	0	3	1	0	0
11	12	50	11	0	1	0	0	0

I have brought the screenshot of python we have imported pandas and numpy then we imported the dataset so this was the dataset. You can see that there are different variables academic ability, parent education, student motivation, advisory evaluation, religion and gender. So I have imported the dataset the data which have stored in this file name academic ability.data. csv file. So this was the output of the data.

(Refer Slide Time: 05:05)

Hypothesis

- Test the hypothesis that “gender and student motivation” are independent

Now we are going to have the hypothesis. Test the hypothesis that gender and student motivations are independent. Now we are going to see is there any connection between the level of motivation and their gender. In many case we presume that the girls students are highly motivated than the boy students. It is in perception that (()) (05:31) whether is there any connection between any gender and the motivation.

So the null hypothesis is that the gender and student motivations are independent. Alternative hypothesis is it is not independent. The hypothesis which we are going to test is gender and student motivations are independent that is our null hypothesis.

(Refer Slide Time: 05:54)

Python code

```
In [19]: #Cross table between gender and student's motivation
obs =pd.pivot_table(acad[['g','sm']],index = 'g',columns='sm',aggfunc=len)
obs
```

```
Out[19]:
```

	sm	0	1	2
0		10	13	6
1		4	9	8

This is very important command for forming our contingency table otherwise call it as a cross table between gender and student motivation. So you see that pd.pivot underscore table the file name acad, g is the gender, sm student motivation, index g should appear in row column should be the sm so aggregate function is length. So after you run this we are getting a contingency table. In contingency table here what is there it is a gender. So in column it is a student motivation.

(Refer Slide Time: 06:35)

Observed values

Gender	Student motivation			Row Sum
	0 (Disagree)	1 (Not decided)	2 (Agree)	
0 (Male)	10	13	6	<u>29</u>
1 (Female)	4	9	8	<u>21</u>
Column Sum	<u>14</u>	<u>22</u>	<u>14</u>	50

In the previous table which are shown in here just I have wrote in the presentation. In the row say 0 means it is male 1 means it is a female that is a code which I have used for gender. The student motivation there are 3 level one is 0 disagree. The question which I have asked is are you willing to spend extra time for getting more marks 0 disagree, 1 not decided, 2 agree. So I have seen the row sum there are 29 male students, there are 21 female students.

So 14 students have told that they will not spend extra time, 22 people have decided that told that they have not decided whether they should spend the extra time to study or not. So 14 people have agreed that they will study, they will spend extra time. Now you see that in row there is a one variable in column there is another variable. The null hypothesis is the gender and the student motivations are independent. We know that the cell the 10, 13, 6 represents the observed frequency. Now we have to find out the expected frequency for each cell.

(Refer Slide Time: 07:54)

Expected frequency (contingency table)

Gender	Student motivation		
	0	1	2
0	$29 \cdot 14 / 50 =$ 8.12	12.76	8.12
1	5.88	9.24	5.88

The expected frequency is as I told you in the previous class that row total multiplied by column total divided by overall. See previous one 29 multiplied by 14 divided by 50. So this 8.12 is the expected frequency. Similarly, for this also we got 12.76, 8.12, 5.88, 9.24, 5.88. One important assumption that the value of the expected frequency should be more than 5. In all the cells the expected frequency is more than 5 then we can continue for our calculation.

(Refer Slide Time: 08:39)

Frequency Table

Gender	Student motivation		
	0	1	2
0	$f_o = 10$ $f_e = 8.12$	$f_o = 13$ $f_e = 12.76$	$f_o = 6$ $f_e = 8.12$
1	$f_o = 4$ $f_e = 5.88$	$f_o = 9$ $f_e = 9.24$	$f_o = 8$ $f_e = 5.88$

So I have wrote f_o represents observed frequency, f_e represents expected frequency. For each cell we know that what is the observed frequency and expected frequency.

(Refer Slide Time: 08:51)

Chi sq. calculation

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

$$= 0.435 + 0.005 + 0.554 + 0.601 + 0.006 + 0.764$$

$$= 2.365$$

Now we have to go for chi-square calculation. We have seen that formula chi-square is calculated value is nothing but observed frequency minus expected frequency whole square divided by expected frequency. So when I go back the first cell observed frequency is 10 – 8.12 whole square divided by 8.12 we will get 0.435. Like this when you do for all the cells and sum it the calculated chi-square value is 2.365.

(Refer Slide Time: 09:26)

Python code

```
In [11]: ## Perform chi2 test to check independence  
         from scipy.stats import chi2_contingency
```

```
In [14]: chi2, p, dof, tbl = chi2_contingency(obs)
```

```
In [15]: chi2
```

Out[15]: 2.3649585225939904

```
In [16]: p
```

Out[16]: 0.3065178579178871

```
In [17]: dof
```

Out[17]: 2

→ Degrees of freedom = $(2-1)*(3-1)$

The python code shows chi-square calculated values for that purpose we have to import chi-square underscore contingency. You see that chi-square, p, degrees of freedom, tbl = chi-square underscore contingency the observed values then when you type chi-square you are getting 2.364 the p value is 0.30. Since the p value is more than 0.05 we have accept null hypothesis.

When we say accept null hypothesis we are concluding that the gender and the level of motivations are independent then you can get the degrees of freedom values also. As I told you the degrees of freedom is number of row - 1, 2 - 1 there are 3 columns 3 - 1 say 2 let us say 2 value. This is the degrees of freedom.

(Refer Slide Time: 10:20)

Python code

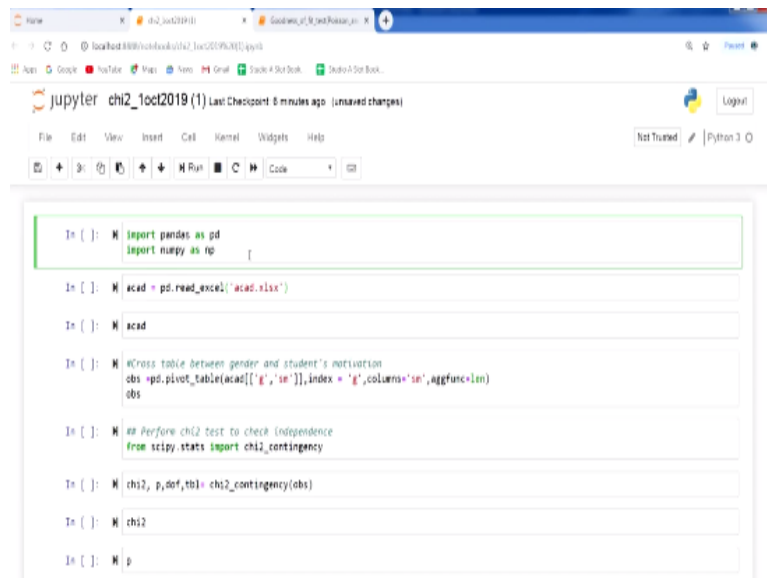
```
In [12]: tbl
```

```
Out[12]: array([[ 8.12, 12.76,  8.12],  
               [ 5.88,  9.24,  5.88]])
```

Contingency table

Then this is contingency table where you can get the expected frequency when you type the tbl you can get the expected frequency you see that this 8.12 this was the I will go back see 8.12, 12.76, 8.12 that value which we got it manually we can directly can get with the help of python. So far I have shown the screenshot of our python programming and explained how to do, how to form a contingency table then how to do the chi-square test. Now I will go to python prompt there I will explain how to input the data and how to do the chi-square test.

(Refer Slide Time: 11:05)



```
In [ ]: import pandas as pd  
import numpy as np
```

```
In [ ]: acad = pd.read_excel('acad.xlsx')
```

```
In [ ]: acad
```

```
In [ ]: #Cross table between gender and student's motivation  
obs = pd.pivot_table(acad[['g', 'se']], index = 'g', columns='se', aggfunc=len)  
obs
```

```
In [ ]: ## Perform chi2 test to check independence  
from scipy.stats import chi2_contingency
```

```
In [ ]: chi2, p, dof, tbl = chi2_contingency(obs)
```

```
In [ ]: chi2
```

```
In [ ]: p
```

I am going to explain how to form a contingency table from the excel file then doing the chi-square test import pandas as pd, import numpy as np I have imported, I have imported, I have stored my dataset and the file name called acad.xlsx. So first I have to run the (()) (11:24).

(Refer Slide Time: 11:25)

```

In [3]: acad = pd.read_excel('acad.xlsx')

In [4]: acad

Out[4]:
   RepNo  aa  pe  sm  r  g  c
0      1  99  10   1  2  0  1
1      2  45  12   0  0  0  0
2      3  57  15   1  1  0  0
3      4  94  18   2  2  1  1
4      5  82  13   2  1  1  1
5      6  59  12   0  0  2  0
6      7  61  12   1  2  0  0
7      8  29   9   0  0  1  1
8      9  35  13   1  1  0  0
9     10  91  16   2  2  1  1
10     11  55  10   0  0  1  0
11     12  58  11   0  1  0  0

```

Now let us see what is the dataset that dataset you see that there are 49 + 0 there are 50 dataset is there, that is the respondent number, academic ability, parent education, student motivation, advisory evaluation, religion, gender and community. We are not going to consider all the variables for our calculation. We are just going to consider only the gender and the student motivation okay.

(Refer Slide Time: 11:53)

```

In [ ]: #Cross table between gender and student's motivation
obs = pd.pivot_table(acad[['g','sm']], index = 'g', columns='sm', aggfunc=len)
obs

In [ ]: # Perform chi2 test to check independence
from scipy.stats import chi2_contingency

In [ ]: chi2, p, dof, tbl = chi2_contingency(obs)

In [ ]: chi2

In [ ]: p

In [ ]: dof

In [ ]: tbl

In [ ]:

```

Then we will form a contingency table for that. This is `obs = pd.pivot` underscore table academic ability that is the file name, column `g` and student motivation and index in row I need to have the `g` value that is the gender value, in column I need to have the student motivation value. So when I see this dataset you see that the output shows that directly I am getting contingency table.

Because when I am explaining theory what happened the contingency table is given to you, but many times that is not the case. The data maybe in some other format you have to create a contingency table before doing the chi-square test. So this command is helping in python, this command is helping us to form the contingency table and it saves lot of our time. So this was the contingency table.

Thus the value in the cell represents the observed frequency. So what this 10 represents when the student level of motivation is 0 the 0 represents male. This is the 10 is our observed value then import chi-square contingency library then we will write chi-square, p, degrees of freedom, tbl = chi-square contingency this obs this obs is wherever contingency table is stored.

So when you run this now the contingency table is run now we want to know the chi-square value the chi-square value is 2.36. This was the chi-square calculated value then we can know the p value, the p value is 0.30 look at the p value which is more than 0.05. So we have to accept our null hypothesis when I say accepting null hypothesis I am concluding that the gender and the student motivations are independent.

There is no connection between the gender and the level of motivation for the student. So we can get to know the degrees of freedom also directly with the help of this command dof and then tbl this give you your expected frequency. If you are doing manually you can compare this expected frequency. So this was the answer which I have shown in my presentation.

(Refer Slide Time: 14:13)

χ^2 Goodness of Fit Test

So far we have seen the first application of chi-square distribution that is test of independency. Now we are moving into another application that is testing, goodness of fit. What is the meaning of goodness of fit. Many time when we collect the data we have to know what distribution this data follows. So the chi-square test is helping us to find out to know what is the distribution this data follows. First you have to take an example of Poisson data then we will check this whether this data follow Poisson distribution or not.

(Refer Slide Time: 14:48)

χ^2 Goodness-of-Fit Test

- The χ^2 goodness-of-fit test compares *expected* (theoretical) *frequencies* of categories from a population distribution to the *observed* (actual) *frequencies* from a distribution to determine whether there is a difference between what was expected and what was observed

What is the chi-square goodness of fit test? Chi-square goodness of fit test compares expected frequencies of categories from population distribution to the observed frequencies, from the distribution to determine whether there is a difference between what was expected and what was observed. So what we are going to do as usual we are also (()) (15:11) to see expected frequencies and observed frequencies. We are going to see is there any difference is there or not.

(Refer Slide Time: 15:16)

χ^2 Goodness-of-Fit Test

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

$$df = k - 1 - p$$

where: f_o = frequency of observed values

f_e = frequency of expected values

k = number of categories

p = number of parameters estimated from the sample data

The formula is same the chi-square value is observed frequency – expected frequency whole square divided by expected frequency. The degrees of freedom this is different from previously what we have seen. In contingency table, the degrees of freedom is number of rows – number of column. Here the degrees of freedom is $k - 1 - p$. The degrees of freedom is $k - 1 - p$ where the k is number of categories number of categories.

Number of observations we can say number of categories here p is number of parameter estimated from the sample data. This the number of parameters you should know in advance. For example, if it is a uniform distribution parameter is 0 if it is a Poisson distribution the parameter is 1 that is only the lambda value. If it is a normal distribution the parameter is 2 because normal distribution is having 2 parameter one is mean and variance.

(Refer Slide Time: 16:16)

Goodness of Fit Test: Poisson Distribution

1. Set up the null and alternative hypotheses.

H_0 : Population has a Poisson probability distribution

H_a : Population does not have a Poisson distribution

2. Select a random sample and

- Record the observed frequency f_i for each value of the Poisson random variable.
- Compute the mean number of occurrences μ .

3. Compute the expected frequency of occurrences e_i for each value of the Poisson random variable.

Now let us follow some steps to test the goodness of fit (GOF) (16:20) given dataset. Now assume that some dataset is given to you we are going to test whether this dataset follows Poisson distribution or not. The first step is to set up the null and alternative hypothesis. What is a null hypothesis? Population has a Poisson probability distribution. What is the alternative hypothesis? Population does not have the Poisson distribution.

Here one important point you have to see so far whenever we see the null hypothesis we say that then the term 'not' will appear in the null hypothesis, but only in the goodness of fit test it is reverse. You see that the given data follow Poisson distribution that should be our null hypothesis. Alternative hypothesis is the data the population does not have a Poisson distribution it is just reverse of that.

For all kinds of hypothesis or testing the word 'not' will appear in the null hypothesis only for the goodness of fit test the word 'not' will appear in your alternative hypothesis. This is one important difference that you have to remember. Select a random sample and record the observed frequency we call as f_i from each value of the Poisson random variable. Compute the mean number of occurrences μ . Because we should know the parameter of Poisson distribution μ compute the expected frequency of occurrences that is e_i for each value of the Poisson random variable.

(Refer Slide Time: 17:58)

Goodness of Fit Test: Poisson Distribution

4. Compute the value of the test statistic

$$\chi^2 = \sum_{i=1}^k \frac{(f_i - e_i)^2}{e_i}$$

where:

f_i = observed frequency for category i

e_i = expected frequency for category i

k = number of categories

Then compute the value of test statistics this is as usual observed frequency – expected frequency whole square divided by expected frequencies. Remember here also the value of expected frequency should be 5 or more. If the expected frequency is not 5 or more you have

to collapse certain intervals and you have to make it so that the expected frequency is 5 or more. We will see that example also here.

(Refer Slide Time: 18:27)

Goodness of Fit Test: Poisson Distribution

5. Rejection rule:

p-value approach: Reject H_0 if $p\text{-value} \leq \alpha$

Critical value approach: Reject H_0 if $\chi^2 \geq \chi_\alpha^2$

where α is the significance level and
there are $k-2$ degrees of freedom

$$k-1-p$$

There are 2 way for rejection rule p-value approach. Reject H_0 the p value is $<$ or $=$ alpha if you follow a critical value approach reject H_0 if the chi-square calculated value is greater than your chi-square critical value which you got from the table where the alpha is significance level and there are $k - 2$ degrees of freedom you should remember how this k because it has come $k - 1 - p$. Because Poisson distribution having one parameter only mean is the parameter for the Poisson distribution. So the value of p is 1 so it has become $k - 2$.

(Refer Slide Time: 19:11)

Goodness of Fit Test: Poisson Distribution

- Example: Parking Garage

In studying the need for an additional entrance to a city parking garage, a consultant has recommended an analysis, that approach is applicable only in situations where the number of cars entering during a specified time period follows a Poisson distribution.

So we will take an example see Parking Garage example. In studying need for an additional entrance to a city parking garage, a consultant has recommended an analysis, consultant has

given some solution, that approach is applicable only in situations where the number of cars entering during a specified time period follows Poisson distribution. Since the consultant has given some solution that can be implemented only if the arrival follow Poisson distribution.

(Refer Slide Time: 19:48)

Goodness of Fit Test: Poisson Distribution

A random sample of 100 one- minute time intervals resulted in the customer arrivals listed below. A statistical test must be conducted to see if the assumption of a Poisson distribution is reasonable.

# Arrivals	0	1	2	3	4	5	6	7	8	9	10	11	12
Frequency	0	1	4	10	14	20	12	12	9	8	6	3	1

A random sample of 100 one minute time interval resulted in customer arrival listed below. A statistical test must be conducted to see if the assumption of a Poisson distribution is reasonable. So what is given a number of arrival is given, the frequency is given so 0 arrival the frequency is 0, 1 arrival the frequency is 1, 2 arrival the frequency is 4, 3 arrival frequency is 10 like that up to 12 arrivals is given.

(Refer Slide Time: 20:20)

Goodness of Fit Test: Poisson Distribution

- Hypotheses

H_0 : Number of cars entering the garage during a one-minute interval is Poisson distributed

H_a : Number of cars entering the garage during a one-minute interval is not Poisson distributed

We will form the hypothesis. What is a null hypothesis number of cars entering the garage during one minute interval is Poisson distributed. Alternative hypothesis is number of cars

entering the garage is during a one minute interval is not Poisson distributed. You see that this is different from our traditional way of making null hypothesis. Generally, the term not will be there in the null hypothesis. But here where the goodness of fit test the term the word not will appear in our alternative hypothesis.

(Refer Slide Time: 20:57)

Python Code

```

In [1]: import scipy
        from scipy.stats import chi2
        from scipy.stats import poisson

In [2]: import pandas as pd
        import numpy as np

In [3]: data = pd.read_excel('P_distribution.xlsx')
        data
Out[3]:
   Arrivals  Frequency
0         0           0
1         1           1
2         2           4
3         3          10
4         4          16
5         5          20
6         6          12
7         7          10
8         8           9
9         9           8
10        10           6
11        11           3
12        12           1

```

This is the python code which I have taken the screenshot import scipy, import chi-square, import Poisson. The dataset I have kept in the file name called P underscore distribution. This was the arrival this is the actual frequency otherwise we can call it as observed frequency.

(Refer Slide Time: 21:21)

Goodness of Fit Test: Poisson Distribution

- Estimate of Poisson Probability Function

$$\text{Total Arrivals} = 0(0) + 1(1) + 2(4) + \dots + 12(1) = 600$$

$$\text{Estimate of } \mu = 600/100 = 6$$

$$\text{Total Time Periods} = 100$$

Hence,

$$= \frac{e^{-\mu} \mu^x}{x!}$$

$$f(x) = \frac{6^x e^{-6}}{x!}$$

$$\mu = \frac{\sum f_n}{\sum f} = \frac{600}{100} = 6$$

The next term we should know mean of the dataset. Estimate of Poisson probability function see the total arrival the mean formula we know that the same simple formula mean is mean $\mu = \frac{\sum fn}{\sum f}$. So f is the frequency n is the number of arrival so 0 into 0 + 1 into 1

like that there will be a value will be 600. So sigma f when I go back you see that this frequency when you add this frequency when you add this frequency that will be 100.

So that value is 6 so the 6 is your mean. We know that our formula traditionally what is our formula. So the formula is $e^{-\mu} \frac{\mu^x}{x!}$ (22:25) $\lambda e^{-\lambda} \frac{\lambda^x}{x!}$. So mu is 6 so 6 to the power x e to the power $-6/x!$.

(Refer Slide Time: 22:44)

Goodness of Fit Test: Poisson Distribution

- Expected Frequencies

x	f(x)	nf(x)	x	f(x)	nf(x)
0	.0025	.25	7	.1377	13.77
1	.0149	1.49	8	.1033	10.33
2	.0446	4.46	9	.0688	6.88
3	.0892	8.92	10	.0413	4.13
4	.1339	13.39	11	.0225	2.25
5	.1606	16.06	12+	.0201	2.01
6	.1606	16.06	Total	1.0000	100.00

Now what we have to do we have to substitute these x values then if you substitute this x values you will get a theoretical frequency. So when $x = 0$ $f(x)$ is when you substitute in this equation when $x = 0$ $6^0 e^{-6} / 0!$ that is 1. So e^{-6} is this 0.0025 this is probability value. We want to know in terms of frequency so that has to be multiple by n n is 100.

When you substitute $x = 1$ in this equation $6^1 e^{-6} / 1!$ will get 0.0149 then multiple by n. So this value will give you the theoretical frequency of Poisson distribution like that we have to have up to 12. You see that here, here the theoretical frequency when you look at the 0, 1, 2 this values. See that this is less than 5 so this has to be added this also less than 5 so these 3 groups has to be grouped.

The same way you see that this is 6.8 here what is happening these values are less than 5. So these values has to be clubbed so that the expected frequency is 5 that is what we have done that one.

(Refer Slide Time: 24:15)

Python code

```
In [4]: Observed_Freq = data['Frequency']

In [5]: total_arrival = 0.00
total_time_period = 100
mu = total_arrival/total_time_period

In [6]: Expected_Freq = []
for i in range(len(Observed_Freq)):
    E_Freq = 100*poisson.pmf(i, mu)
    Expected_Freq.append(E_Freq)

In [7]: Expected_Freq

Out[7]: [0.2478752176661584,
1.4872511059498185,
4.461753937939446,
8.923507875878892,
13.385261753408332,
16.862314184793995,
16.862314184793995,
13.767097884112569,
10.1757735104447,
6.88384892095786,
4.110709141233756,
2.252866831091267,
1.126448802546081]
```

Okay so we have got the observed frequency we got that mean value then for each x value we got our expected frequency value by using the (()) (24:25) okay.

(Refer Slide Time: 24:28)

Python code

```
In [4]: expected_freq_round_off = [round(x, 2) for x in expected_freq]
expected_freq_round_off

Out[4]: [0.25,
1.49,
4.46,
8.92,
13.39,
16.86,
16.86,
13.77,
10.18,
6.88,
4.11,
2.25,
1.13]

In [4]: df = pd.DataFrame([Observed_Freq, expected_freq_round_off ], columns = ['Observed Frequency', 'Expected Frequency'])
df

Out[4]:
```

	Observed Frequency	Expected Frequency
0	0	0.25
1	1	1.49
2	4	4.46
3	10	8.92
4	14	13.39
5	20	16.86
6	12	16.86
7	6	13.77
8	3	10.18
9	1	6.88
10	0	4.11
11	1	2.25
12	1	1.13

So then we will do round off this was our rounded value using python. When you look at this 0.25 you go back 0.25, 1.49 you will get exact the same value. Now we are going to have only 2 column one is observed frequency next one is expected frequency.

(Refer Slide Time: 24:48)

Goodness of Fit Test: Poisson Distribution

- Observed and Expected Frequencies

i	f_i	e_i	$f_i - e_i$
<u>0 or 1 or 2</u>	<u>5</u>	<u>6.20</u>	-1.20
3	10	8.92	1.08
4	14	13.39	0.61
5	20	16.06	3.94
6	12	16.06	-4.06
7	12	13.77	-1.77
8	9	10.33	-1.33
9	8	6.88	1.12
10 or more	10	<u>8.39</u>	1.61

You see that 0 or 1 or 2 so that are clubbed so that the expected frequency is more than 5. Similarly, 10, 11, 12 these are grouped together so that the expected frequency is 8.39 otherwise it will be less than 5. Now how many numbers of interval is there 1, 2, 3, 4, 5, 6, 7, 8, 9 interval is there.

(Refer Slide Time: 25:15)

Python code

```
In [10]: obs_freq = [5, 10, 14, 20, 12, 12, 9, 8, 10]
         expected_freq = [6.20, 8.92, 13.39, 16.06, 16.06, 13.77, 10.33, 6.88, 8.39]

In [11]: scipy.stats.chisquare(obs_freq, expected_freq)

Out[11]: Power_divergenceResult(statistic=3.2738182931105193, pvalue=0.916017731732134)
```

Now this is observed frequency expected frequency. Now if you directly you can run this command that is `scipy.stats.chisquare` observed frequency – expected frequency that we are getting 3.27 the p value is 0.911 that is more than 0.05. So we have to accept null hypothesis when we accept null hypothesis we are concluding that the given arrival pattern follow Poisson distribution.

(Refer Slide Time: 25:45)

Goodness of Fit Test: Poisson Distribution

- **Rejection Rule**

With $\alpha = .05$ and $k - p - 1 = 9 - 1 - 1 = 7$ d.f.

(where k = number of categories and p = number of population parameters estimated),

$$\chi^2_{.05} = 14.067$$

Reject H_0 if $p\text{-value} \leq .05$ or $\chi^2 \geq 14.067$.



- **Test Statistic**

$$\chi^2 = \frac{(-1.20)^2}{6.20} + \frac{(1.08)^2}{8.92} + \dots + \frac{(1.61)^2}{8.39} = 3.268$$

You see that for rejection rule when $\alpha = 0.05$ we have got $k = 9$ as I told you because 9 interval and going back I will explain 1, 2, 3, 4, 5, 6, 7, 8, 9 interval that is why here $k = 9$. p is the number of parameter as we know that that Poisson distribution having only one parameter so it is $9 - 1 - 1$ so 7 degrees of freedom. For 7 degrees of freedom when $\alpha = 0.05$ we can get the chi-square table value is 14.06.

See when you look at the chi-square calculated values it is 3.268. So this value will lie on the acceptance side. So we have to accept because the chi-square distribution to be like this so this is 14.067 you are 3.268 will be here it will be lying on the acceptance side so you have to accept the null hypothesis.

(Refer Slide Time: 26:44)

Python code

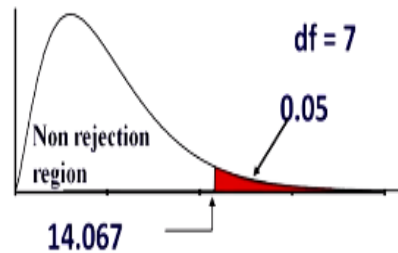
```
In [4]: from scipy.stats import chi2
        chi2.ppf(0.95,7)
```

```
Out[4]: 14.067140449340167
```

The same thing 0.957 so the chi-square calculated value is 14.06.

(Refer Slide Time: 26:52)

Goodness of Fit Test: Poisson Distribution



$$\chi^2_{Cal} = \underline{3.268} < 14.067, \text{ do not reject } H_0.$$

See there 14.06 but not calculated value. The chi-square table value is 14.06, but our calculated value is 3.268 so it is lying on the acceptance side do not reject null hypothesis then we will conclude that the arrival pattern follow Poisson distribution. In this class, I have explained how to form a contingency table after forming contingency table how to do the chi-square test with the help of python.

The next topic which I have started testing goodness of fit. Suppose some dataset is given if you want to test what distribution it follows. I have taken some dataset then I have tested whether this dataset follow Poisson distribution or not. I have explained the python screenshot. In the next class I will run the python code for testing Poisson distribution for given dataset and I will explain how to test goodness of fit for uniform distribution and normal distribution that we will see in the next class. Thank you.