Department of AI & DS
CSE and CS&IT

COURSE NAME: PROBABILITY, STATISTICS AND QUEUING THEORY

COURSE CODE: 23MT2005

Topic

Chi square test

Session - 18

AIM OF THE SESSION

To familiarize students with the Test of good ness of fit and independence of attributes using Chi square test

INSTRUCTIONAL OBJECTIVES

This Session is designed to:

1. Define Chisqare test
2. Check the conditions for the validity of chi-square test
3. Test statistic for goodness of fit and independence of attributes

LEARNING OUTCOMES

At the end of this session, you should be able to:

1. Define Null and alternative hypothesis of test of significance using chisquare
2. Describe the procedure for Chi square test
3. Summarize the importance of Chisquare for testing goodness of fit and independence of attributes

Chi-Square test

The Chi-square (χ^2) test measures the alignment between two sets of frequency measures. It performs two types of functions namely

(i) Goodness of fit: A common use is to assess whether a measure/ observed set of measures follows an expected pattern. The expected frequency may be determined from prior knowledge or by calculation of an average from the given data.

The null hypothesis, H_0 is that the two sets of measures are not significantly different.

(i) Measures of Independence: The chi-square test can be used in the reverse manner to goodness of fit. If the two sets of measures are compared, then just as you can show they align, you can also determine if they do not align.

The null hypothesis here is that the two sets of measures are similar.

Chi-Square test

The experimental hypothesis evaluated with the chi-square goodness of fit test is whether or not there is a difference between the observed frequencies of the k cells and their expected frequencies. The expected frequency of a cell is determined through the use of probability theory or is based on some pre existing empirical information about the variable under study. If the result of the chi-square goodness of fit test is significant, the researcher can conclude that in the underlying population represented by the sample there is a high likelihood that the observed frequency for at least one of the k cells is not equal to the expected frequency of the cell.

Symbolically,

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i} \sim \chi^2{}_{(n-1)}\text{d.f.}$$

Chi-Square test

Chi-square test is performed which measures the discrepancy between the observed frequencies and theoretically determined frequencies from the assumed distribution of the same event.

Conditions for the validity of the χ^2- test: χ^2 - test is an approximate test for large values of n. For the validity of chi-square test of ‘goodness of fit’ between theory and experiment, the following conditions must be satisfied:

(i) The sample observations should be independent

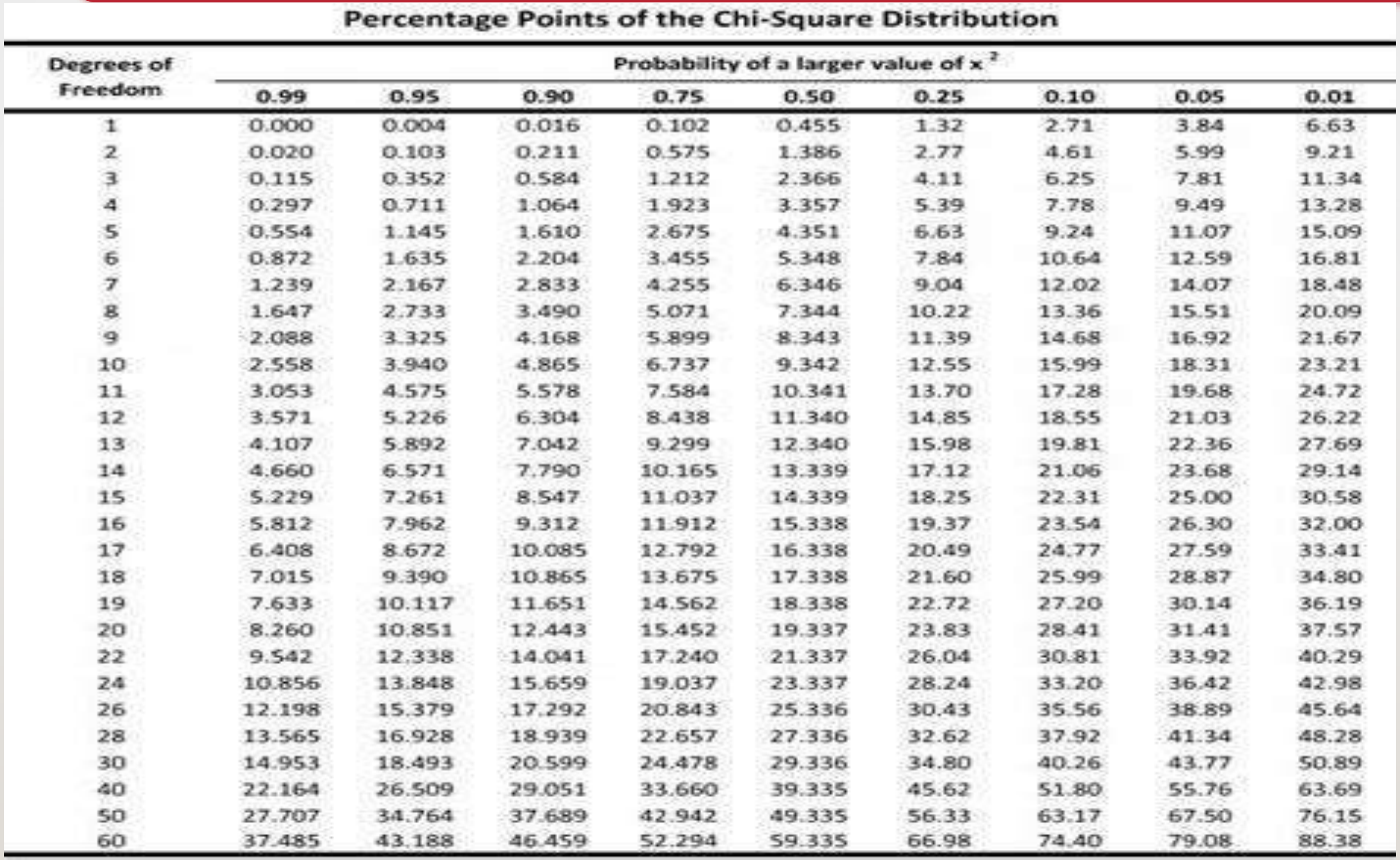
(ii) Constraints on the cell frequencies, if any, should be linear.

(iii) N, the total frequency should be reasonably large, say greater than 50.

(iv) No theoretical cell frequency should be less than 5.

- χ^2-test depends only one the set of observed frequencies and expected frequencies and on degrees of freedom. It does not involve any population parameters, it is termed as statistic and the test is known as the Non-parametric test or Distribution free test.

ACTIVITIES/ CASE STUDIES/ IMPORTANT FACTS RELATED TO THE SESSION



Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of x ^{2}								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59	33.41
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87	34.80
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14	36.19
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41	37.57
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92	40.29
24	10.856	13.848	15.659	19.037	23.337	28.24	33.20	36.42	42.98
26	12.198	15.379	17.292	20.843	25.336	30.43	35.56	38.89	45.64
28	13.565	16.928	18.939	22.657	27.336	32.62	37.92	41.34	48.28
30	14.953	18.493	20.599	24.478	29.336	34.80	40.26	43.77	50.89
40	22.164	26.509	29.051	33.660	39.335	45.62	51.80	55.76	63.69
50	27.707	34.764	37.689	42.942	49.335	56.33	63.17	67.50	76.15
60	37.485	43.188	46.459	52.294	59.335	66.98	74.40	79.08	88.38

Example 1

The Acme Battery Company has developed a new cell phone battery. On average, the battery lasts 60 minutes on a single charge. The standard deviation is 4 minutes. Suppose the manufacturing department runs a quality control test. They randomly select 7 batteries. The standard deviation of the selected batteries is 6 minutes. What would be the chi-square statistic represented by this test?

Solution:

Given, N=7 (number of observations).

The standard deviation of the population is 4 minutes.

The standard deviation of the sample is 6 minutes.

To calculate the χ^2 statistic, we use the following chi-square equation

$$\chi^2 = [(n-1) * s^2]/\sigma^2$$

$$\chi^2 = \frac{[(7-1)*6^2]}{4^2} = 13.5$$

where χ^2 is the test statistic, n is the sample size, s is the standard deviation of the sample, σ is the standard deviation of the population.

Example 2:

Example 2: The following table represents the number of boys and the number of girls who choose each of the five possible answers to an item in an attitude scale.

	Approve strongly	Approve	indifferent	disapprove	Strongly disapprove	Total
Boys	25	30	10	25	10	100
Girls	10	15	5	15	15	60
Total	35	45	15	40	25	160

Do these data indicate a significant difference in attitude towards this question? (Note : Test the independence (null hypothesis)

Solution: To examine whether there any significant difference between the boys and girls in attitude towards the question or not

Step 1: Set up the null hypothesis: The attitude towards questions are independent.

Now, under the null hypothesis $H_{0,}$ the test statistic is

Example 2:

$\chi^2 = \sum(\frac{(oi-ei)^2}{ei})$

Where O_i= observed frequencies

e_i = expected frequencies which is given by

expected frequency = (row total X column total) /grand total

The above test statistic χ^2 follows chi-square distribution at (r-1) (s-1) degrees of freedom.

Example 2:

Calculations:

The given data can be tabulated as follows:

	Approve strongly	Approve	indifferent	disapprove	Strongly disapprove	Total
Boys	25	30	10	25	10	100
Girls	10	15	5	15	15	60
Total	35	45	15	40	25	160

$$\text{E (25)} = \frac{100X35}{160} = 21.875, \text{E (30)} = \frac{100X45}{160} = 28.125, \text{E (10)} = \frac{100X15}{160} = 9.375, \text{E (25)} = \frac{100X40}{160} = 25$$

$$\text{E (10)} = \frac{100X25}{160} = 15.625,\ \text{E (10)} = \frac{60X35}{160} = 13.125, \text{E (15)} = \frac{60X45}{160} = 16.875, \text{E (5)} = \frac{60X15}{160} = 5.625$$

$$\text{E (15)} = \frac{60X40}{160} = 15, \text{E (15)} = \frac{60X25}{160} = 9.375$$

Example 2:

Observed frequency (o_i)	Expected frequency (e_i)	o_i-e_i	$(oi - ei)^2$	$\frac{(o_i - ei)^2}{ei}$
25	21.875	3.125	9.7656	0.4464
30	28.125	1.875	3.5156	0.1250
10	9.375	0.625	0.3906	0.0417
25	25	0	0	0
10	15.625	-5.625	31.6406	2.0250
10	13.125	-3.125	9.7656	0.7440
15	16.875	-1.875	3.5156	0.2083
5	5.625	-0.625	0.3906	0.0694
15	15	0	0	0
15	9.375	5.625	31.6406	3.3750
160	160			7.0348

Now, under the null hypothesis H_0, the test statistic is

$$\chi^2 = \sum\left(\frac{(oi - ei)^2}{ei}\right) = 7.0348$$

The table value of χ^2 at (r-1)(s-1) = (5-1)(2-1) =4d.f. and at 5% level of significance is 9.49.

χ^2 calculated value is less than χ^2 table value, so we accept the null hypothesis H_0.

Hence, we may conclude that there is no significant difference towards the question between the boys and girls that they are independent.

DEEMED TO BE UNIVERSITY

SUMMARY

In this session, the concept of Chi square test for goodness of fit and independence of attributes

1. Define Chi square test
2. Discuss in detail about conditions for the validity of chi square distribution

SELF-ASSESSMENT QUESTIONS

Which of these distributions is used for a testing hypothesis?

a) Normal Distribution
b) Chi-Squared Distribution
c) Gamma Distribution
d) Poisson Distribution

The Chi-square test is primarily used for which type of data?

A. Interval data
B: Continuous data
C: Nominal or categorical data
D: Ordinal data

TERMINAL QUESTIONS

1) A business owner had been working to improve employee relations in his company. He predicted that he met his goal of increasing employee satisfaction from 65% to 80%. Employees from four departments were asked if they were satisfied with the working conditions of the company. The results are shown in the following table:

	Finance	Sales	Human Resources	Technolog y	
Satisfied	12	38	5	8	
Dissatisfied	7	19	3	1	
Total	19	57	8	9	

We can use chi square to determine whether the results support or reject the business owner's prediction.

2. The following table gives the number of accidents that work place in an industry surveying varies days of the week. Test if the accidents are uniformly distributed over the week.

Days:	Mon	Tue	Wed	Thu	Fri	Sat
No.of accidents	14	18	12	11	15	14

TERMINAL QUESTIONS

3. Among 64 off springs of a certain cross between guinea pigs 34 were red, 10 were black and 20 were white. According to the genetic model these numbers should be in the ratio 9:3:4. Are the data consistent with the model at 5% level.

4. The figures given below are (a) the observed frequencies (b) the theoretical frequencies of a normal distribution respectively:

(a) oi	1	5	20	28	42	22	15	5	2
(b)ei	1	6	18	25	40	25	18	6	1

Applying the - test of goodness of fit for the above data and comment?

REFERENCES FOR FURTHER LEARNING OF THE SESSION

Reference Books:

1. William Feller, An Introduction to Probability Theory and Its Applications: Volime 1, Third Edition, 1968 by John Wiley & Sons,Inc.
2. Alex Tsun, Probability & Statistics with Applications to Computing (Available at: http://www.alextsun.com/files/Prob_Stat_for_CS_Book.pdf)

3. Richard A Johnson, Miller& Freund's Probability and statistics for Engineers, PHI, New Delhi, 11th Edition (2011).

Sites and Web links:

1. https://www.khanacademy.org/math/statistics-probability/significance-tests-one-sample/more-significance-testing-videos/v/small-sample-hypothesis-test

THANK YOU

Team – PSQT EVEN SEMESTER 2024-25