IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

χ^2 Test of Independence - I

Dr. A. Ramesh

DEPARTMENT OF MANAGEMENT STUDIES

Agenda

- To understand χ^2 Test of Independence

$\bar{x} \to \mu$

1 sample Z-test

1 sample Z proportion test

$\bar{x}_1$ $\bar{x}_2$

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$

two sample Z

t

2 sample Z-proportion test

$H_0: \mu_1 = \mu_2 = \mu_3$

$H_1: \mu_1 \neq \mu_2 \neq \mu_3$

ANOVA

Chi-squared test

χ^2 Test of Independence

- The chi-square test of independence is a statistical method used to determine whether there is a significant association between two categorical variables.
- It assesses whether the observed frequencies of the categories in one variable are dependent on the categories of the other variable or if they occur independently.
- Qualitative Variables
- Nominal Data

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

Contingency Table

Geographic Region × Type of financial Investment

	E	F	G	
A			O_{13}	n_A
B				n_B
C				n_C
D				n_D
	n_E	n_F	n_G	N

χ^2 Test of Independence: Investment Example

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

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

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

$$e_{AF} = N \cdot P(A \cap F)$$
$$= \cancel{N}\left(\frac{n_A}{\cancel{N}} \cdot \frac{n_F}{N}\right)$$
$$= \frac{n_A \cdot n_F}{N}$$

Contingency Table

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

χ^2 Test of Independence: Formulas

Expected Frequencies

$$e_{ij} = \frac{(n_i)(n_j)}{N}$$

where: i = the row

j = the column

n_i = the total of row i

n_j = the total of column j

N = the total of all frequencies

χ^2 Test of Independence: Formulas

Calculated χ^2
(Observed χ^2)

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

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

Example for Independence

χ^2 Test of Independence

H_o : Type of gasoline is independent of income

H_a : Type of gasoline is not independent of income

χ^2 Test of Independence

r = 4

c = 3

Income	Type of Gasoline		
	Regular	Premium	Extra Premium
Less than \$30,000			
\$30,000 to \$49,999			
\$50,000 to \$99,000			
At least \$100,000			

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

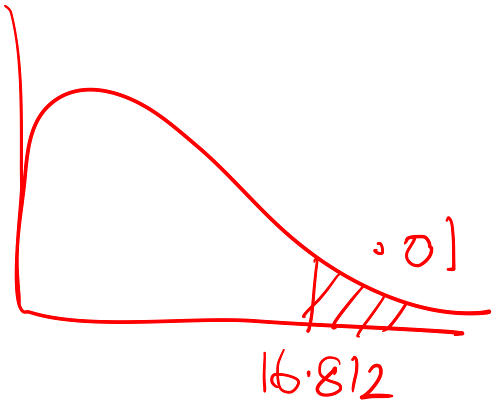
$$\alpha = .01$$

$$df = (r-1)(c-1)$$

$$= (4-1)(3-1)$$

$$= 6$$

$$\chi^2_{.01,6} = 16.812$$



If $\chi^2_{\text{Cal}} > 16.812$, reject Ho.

If $\chi^2_{\text{Cal}} \le 16.812$, do not reject Ho.

Python code

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

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

Out[6]: 16.811893829770927
```

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

Gasoline Preference Versus Income Category: Expected Frequencies

$$e_{ij} = \frac{(n_i)(n_j)}{N}$$

$$e_{11} = \frac{(107)(238)}{385} = 66.15$$

$$e_{12} = \frac{(107)(88)}{385} = 24.46$$

$$e_{13} = \frac{(107)(59)}{385} = 16.40$$

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

Gasoline Preference Versus Income Category: χ^2 Calculation

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

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

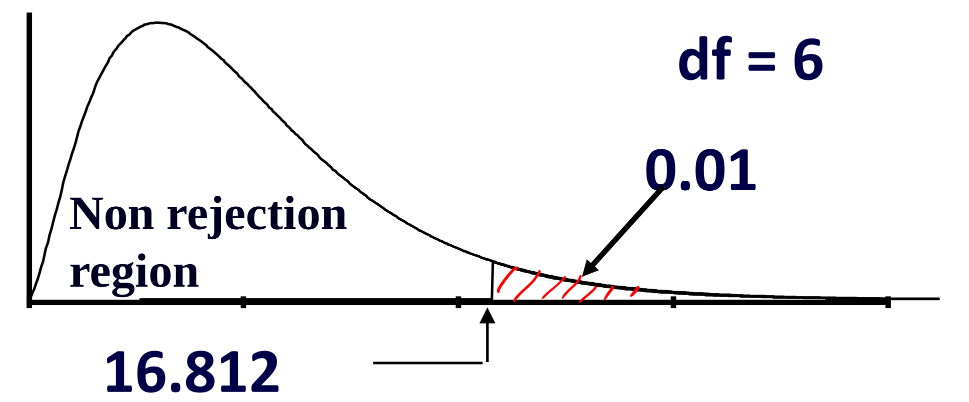
$$\frac{(102-87.78)^2}{87.78} + \frac{(27-32.46)^2}{32.46} + \frac{(13-21.76)^2}{21.76} +$$

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

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

$$= 70.75$$

Gasoline Preference Versus Income Category: Conclusion



$$\chi^2_{Cal} = 70.7\overset{5}{8} > 16.812, \text{ reject } H_o.$$

Contingency Tables

- Contingency tables, also known as cross-tabulations or crosstabs, are a valuable tool in statistics, particularly in the analysis of categorical data.
- Useful in situations involving multiple population proportions
- Used to classify sample observations according to two or more characteristics
- By displaying the frequencies or counts of observations for each combination of categories, they offer a visual representation of the data's structure.

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

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

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.

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

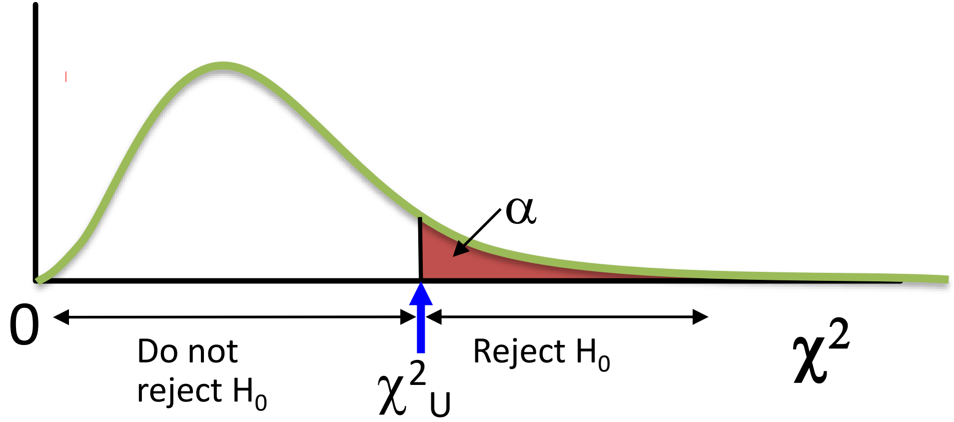
Assumed: each cell in the contingency table has expected frequency of at least 5

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



Observed vs. Expected Frequencies

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

The Chi-Square Test Statistic

Hand Preference	Gender: Female	Gender: 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:

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

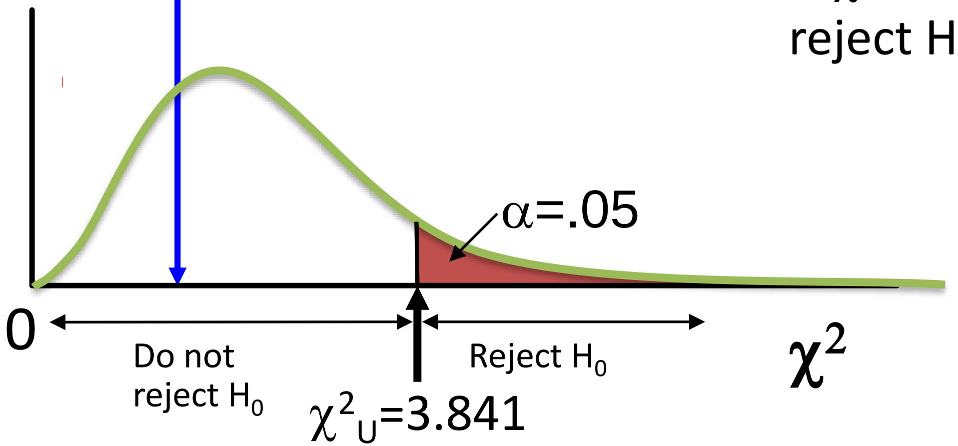
$$= \frac{(12-14.4)^2}{14.4} + \frac{(108-105.6)^2}{105.6} + \frac{(24-21.6)^2}{21.6} + \frac{(156-158.4)^2}{158.4} = \boxed{0.7576}$$

The Chi-Square Test Statistic

The test statistic is $\chi^2 = \boxed{0.7576,}$ χ_U^2 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.

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:

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

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

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 x c table
- f_e = expected frequency in a particular cell if H_0 is true
- χ^2 for the 2 x 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

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:

χ2 Test with More Than Two Proportions: Example

Object to Record Sharing	Organization		
	Insurance Companies π_1	Pharmacies π_2	Medical Researchers π_3
Yes	410	295	335
No	90	205	165

χ2 Test with More Than Two Proportions: Example

Object to Record Sharing	Organization			
	Insurance Companies	Pharmacies	Medical Researchers	Row Sum
Yes	410 $\frac{1040 \times 500}{1500}$	295 $\frac{1040 \times 500}{1500}$	335	1040
No	90	205	165	460
Column Sum	500	500	500	1500

$\chi 2$ Test with More Than Two Proportions: Example

The overall proportion is:

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

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	$f_o = 90$ $f_e = 153.333$	$f_o = 205$ $f_e = 153.333$	$f_o = 165$ $f_e = 153.333$

$\chi 2$ 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:

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

χ^2 Test with More Than Two Proportions: Example

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

H_1: Not all of the π_j are equal (j = 1, 2, 3)

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

χ^2_U = 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