MATHEMATICS FOR COMPUTER SCIENCE ENGINEERS

HYPOTHESIS and INFERENCE

Dr. Deepa Nair
Department of Science and Humanities

MATHEMATICS FOR COMPUTER SCIENCE ENGINEERS

UNIT-3 HYPOTHESIS and INFERENCE

Chi-squared Test

Dr. Deepa Nair

Department of Science and Humanities

Chi-squared Test

- Chi square test is used to **measure independence of two categorical variables**. i.e if one variable has any affect on another
- Ex:Does the gender of the person determine which chocolate they like?

- It is also **used to measure goodness of fit**. i.e if the observed and expected values match.

- For example, your model expected women winning lottery is 0.05 more than male, but is that really the case. Does the real data really match up with your prediction?

- Chi square is used when data is categorical (i.e can be classified into groups (yes/no)(red/blue/green))

PES UNIVERSITY

Tender for collecting toll for a newly opened bridge

Day	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
No	50	20	90	130	200	170	220

Day	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
No	50	50	100	130	200	150	200

You want to know if the day of week affects
The number of people sending tenders for the bridge?

You accordingly set up your H0 and H1

Ex:H0→probability of all day same

If H0 rejected then the day affects number of tenders else not

Chi-squared Test

- A generalization of the Bernoulli trial is the multinomial trial

- Which is an experiment that can result in any one of k outcomes, where $k \geq 2$.

- Theprobabilities of the k outcomes are denoted $p_1, \ldots, p_k$.

- The null hypothesis has the form

$H_0 : p_1 = p_{01}, p_2 = p_{02}, \ldots p_k = p_{0k}$.(these are your expected probabilities)

For example:

A gambler wants to test a die to see whether it deviates from fairness.

Let$p_i$be the probability that the number i comes up. The null hypothesis will state that the die is fair.

The null hypothesis is $H_0 : p_1 = p_{01}, p_2 = p_{02}, \ldots p_6 = p_{06} = 1/6$.

- The gambler rolls the die 600 times and The results obtained are called the observed values.

- To test the null hypothesis, we construct a second column, labeled “Expected.” This column contains the expected values.

- The expected value for a given outcome is the mean number of trials that would result in that outcome if H_0 were true.

- The idea behind the hypothesis test is that if H_0 is true, then the observed and expected values are likely to be close to each other.

- If H0 is false that implies that there is some hidden factor affecting the outcome (i.e the dice is not fair)

- Therefore we will construct a test statistic that measures the closeness of the observed to the expected values.

- The statistic is called the chi-square statistic. To define it, let k be the number of outcomes ($k = 6$ in the die example),
- Let Oi and Ei be the observed and expected numbers of trials, respectively, that result in outcome i .
- The chi-square statistic is

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

Chi-squared Test

- When the expected values are all sufficiently large, a good approximation is available.

- It is called the chi-square distribution with $k - 1$ degrees of freedom, denoted $\chi 2_{k-1}$

- A table for the chi-square distribution is available
- We look up the p-value associated for k-1 degree of freedom and given significance level (alpha) [this is critical value]
- If the calculated $\chi 2$ is greater than critical value →reject null hypothesis
- Else →fail to reject null hypothesis

OR

You can look up p-val associated with chi-square closed to your calculated chi-square and if it is less than significance level→reject null hypothesis

Example:

Consider a die is thrown 600 times with following results.

Number turned up	1	2	3	4	5	6
Frequency	115	97	91	101	110	86

Is the die unbiased at 10% significance level?

Example:

Expected value=600/6=100 (unbiased die means equal outcome of all)

Hypothesis

H0: Die is unbiased

H1: Die is biased

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

Catogory	Observed	Expected
1	115	100
2	97	100
3	91	100
4	101	100
5	110	100
6	86	100
Tot	600	600

Chi-squared Test

$$\chi 2 = \frac{(115-100)^2}{100} + \cdots \frac{(86-100)^2}{100}$$

$$100$$

$$= 2.25 + \cdots + 1.96$$

$$= 6.12$$

Degrees of freedom=n-1=6-1=5

$\chi 2$ at 10% for df 5 is 9.236

Since 6.12<9.236

We fail to reject H0,

Conclude that die is unbiased

	P					
DF	0.995	0.975	0.2	0.1	0.05	0.025
1	.0004	.00016	1.642	2.706	3.841	5.024
2	0.01	0.0506	3.219	4.605	5.991	7.378
3	0.0717	0.216	4.642	6.251	7.815	9.348
4	0.207	0.484	5.989	7.779	9.488	11.143
5	0.412	0.831	7.289	9.236	11.07	12.833
6	0.676	1.237	8.558	10.645	12.592	14.449
7	0.989	1.69	9.803	12.017	14.067	16.013
8	1.344	2.18	11.03	13.362	15.507	17.535
9	1.735	2.7	12.242	14.684	16.919	19.023
10	2.156	3.247	13.442	15.987	18.307	20.483

The Chi-Square Test for Homogeneity:

- Sometimes several multinomial trials are conducted, each with the same set of possible outcomes.

- The null hypothesis is that the probabilities of the outcomes are the same for each experiment.
- We calculate the expected values for each cell using the formula
- $$E = \frac{(\text{Row Total}) \times (\text{Column Total})}{\text{Grand total}}$$
- Then calculate overall chi-square (applying formula over all cells observed and expected)

- Then calculate overall chi-square (applying formula over all cells observed and expected)

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

- Then calculate degree of freedom using formula

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

where r is the number of rows and c is the number of columns.

- Then look up critical X2 val and compare with calculated to determine whether to reject null hypothesis or not

The Chi-Square Test for Homogeneity:

Example:

- Use the following data to test the null hypothesis that the proportions of pins that are too thin, OK, or too thick are the same for all the machines.

The Chi-Square Test for Homogeneity:

Example: specification. A pin may meet the specification, or it may be too thin or too thick. Pins are sampled from each machine, and the number of pins in each category is counted. Table below presents the results. Use the data in Table to test the null hypothesis that the proportions of pins that are too thin, OK, or too thick are the same for all the machines.

	TOO thin	OK	Too thick
Machine 1	10	102	8
Machine 2	34	161	5
Machine 3	12	79	9
Machine 4	10	60	10

Null Hypothesis (H_0): **The proportions of pins that are too thin, OK, or too thick are the same across all machines.**
Alternative Hypothesis (H_1): **The proportions of pins that are too thin, OK, or too thick differ for at least one machine.**

Expected Value Formula

For each cell in the table, the **expected value** is calculated as:

E=(Row Total)×(Column Total)/Grand total

Note: calculate the expected values for each cell. The apply in chi square formula

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

Solution:

Expected values for Table

	Too Thin	OK	Too Thick	Total
Machine 1	15.84	96.48	7.68	120.00
Machine 2	26.40	160.80	12.80	200.00
Machine 3	13.20	80.40	6.40	100.00
Machine 4	10.56	64.32	5.12	80.00
Total	66.00	402.00	32.00	500.00

Observed values

	T00 thin	OK	Too thick
Machine 1	10	102	8
Machine 2	34	161	5
Machine 3	12	79	9
Machine 4	10	60	10

$$\chi 2 = \frac{(10 - 15.84)^2}{15.84} + \cdots + \frac{(10 - 5.12)^2}{5.12}$$

$$= \frac{34.1056}{15.84} + \cdots + \frac{23.8144}{5.12}$$

$$\boldsymbol{\chi 2 = 15.5844}$$

Solution:

- Since there are four rows and three columns, the number of degrees of freedom is $(4-1)(3-1) = 6$.
- To obtain the *P*-value, we consult the chi-square table Looking under six degrees of freedom, we find that the upper 2.5% point is 14.449, and the upper 1% point is 16.812.
- Therefore $0.01 < P < 0.025$. We reject H0 (since no matter what the p-val is less than 0.05) (default alpha is 5%)
- therefore It is reasonable to conclude that the machines differ in the proportions of pins that are too thin, OK, or too thick.

The Chi-Square Test for Independence:

- In some cases, both row and column totals are random. In either case, we can test the null hypothesis that the probabilities of the column outcomes are the same for each row outcome, and the test is exactly the same in both cases.

The Chi-Square Test for Independence:

Example:

The cylindrical steel pins in Example 6.21 are subject to a length specification as well as a diameter specification. With respect to the length, a pin may meet the specification, or it may be too short or too long. A total of 1021 pins are sampled and categorized with respect to both length and diameter specification. The results are presented in the following table. Test the null hypothesis that the proportions of pins that are too thin, OK, or too thick with respect to the diameter specification do not depend on the classification with respect to the length specification.

Observed Values for 1021 Steel Pins

	Diameter			
Length	Too Thin	OK	Too Thick	Total
Too Short	13	117	4	134
OK	62	664	80	806
Too Long	5	68	8	81
Total	80	849	92	1021

Example:

Formulate the Hypotheses:

- **Null Hypothesis (H_0):** The proportions of pins that are too thin, OK, or too thick with respect to the diameter specification do not depend on the length specification.
- **Alternative Hypothesis (H_1):** The proportions of pins that are too thin, OK, or too thick with respect to the diameter specification depend on the length specification.

Calculate the expected values

The expected frequency E_{ij} for each cell under the assumption of independence is calculated using:

$$E_{ij} = \frac{(\text{Row Total}_i \times \text{Col Total}_j)}{\text{Grand Total}}$$

For each combination of row (length category) and column (diameter category), calculate the expected frequencies.

MATHEMATICS FOR COMPUTER SCIENCE ENGINEERS

Chi-squared Test

Expected Values for 1021 Steel Pins

Length	Diameter: Too Thin	OK	Too Thick	Total
Too Short	10.50	111.43	12.07	134.0
OK	63.15	670.22	72.63	806.0
Too Long	6.35	67.36	7.30	81.0
Total	80.0	849.0	92.0	1021.0

Observed Values for 1021 Steel Pins

Length	Diameter: Too Thin	OK	Too Thick	Total
Too Short	13	117	4	134
OK	62	664	80	806
Too Long	5	68	8	81
Total	80	849	92	1021

$$\chi 2 = \frac{(13 - 10.50)^2}{10.50} + \cdots + \frac{(8 - 7.30)^2}{7.30}$$

$$= \frac{6.25}{10.50} + \cdots + \frac{0.49}{7.30} = 7.46$$

Chi-squared Test

- Since there are three rows and three columns, the number of degrees of freedom is $(3 - 1)(3 - 1) = 4$.

- To obtain the *P*-value, we consult the chi-square table Looking under four degrees of freedom, we find that the upper 10% point is 7.779. We conclude that $P > 0.10$.
- We reject null hypothesis

- There is no evidence that the length and thickness are related.

Chi-squared Test-Practice Problems (independence test)

- At an assembly plant for light trucks, routine monitoring of the quality of welds yields the following data:

	Number of Welds		
	High Quality	Moderate Quality	Low Quality
Day Shift	467	191	42
Evening Shift	445	171	34
Night Shift	254	129	17

Can you conclude that the quality varies among shifts?

a. State the appropriate null hypothesis.

b. Compute the expected values under the null hypothesis.

c. Compute the value of the chi-square statistic.

d. Find the P-value. What do you conclude?

Chi-squared Test-Practice Problems (homogeneity test)

- A survey of adults with diabetes. Each respondent was categorized by gender and income level.

Can you conclude that the proportions in the various income categories differ between men and women?

	Poor	Near Poor	Low Income	Middle Income	High Income
Men	156	77	253	513	604
Women	348	152	433	592	511

Dr. Deepa Nair

Department of Science and Humanities

deepanair@pes.edu