Statistics

Chi-Square test

In recent years, the use of specialized statistical methods for categorical data has increased dramatically, particularly for applications in the biomedical and social sciences. Categorical scales occur frequently in the health sciences, for measuring responses. E.g.

- patient survives an operation (yes, no),
- severity of an injury (none, mild, moderate, severe), and
- stage of a disease (initial, advanced).

Studies often collect data on categorical variables that can be summarized as a series of counts and commonly arranged in a tabular format known as a **contingency table.**

Chi-Square test χ^2

The most obvious difference between the chi-square tests and the other hypothesis tests we have considered (T test) is the *nature of the data (categorical data).*

- For chi-square, the data are ***frequencies*** rather than numerical scores.
- Used for testing significance of patterns in qualitative data.
- Test statistic is based on counts (frequencies) that represent the number of items that fall in each category
- Test statistics measures the agreement between actual counts(observed) and expected counts assuming the null hypothesis

Chi-Square test χ^2

Chi-square Test – Distribution table and formulas

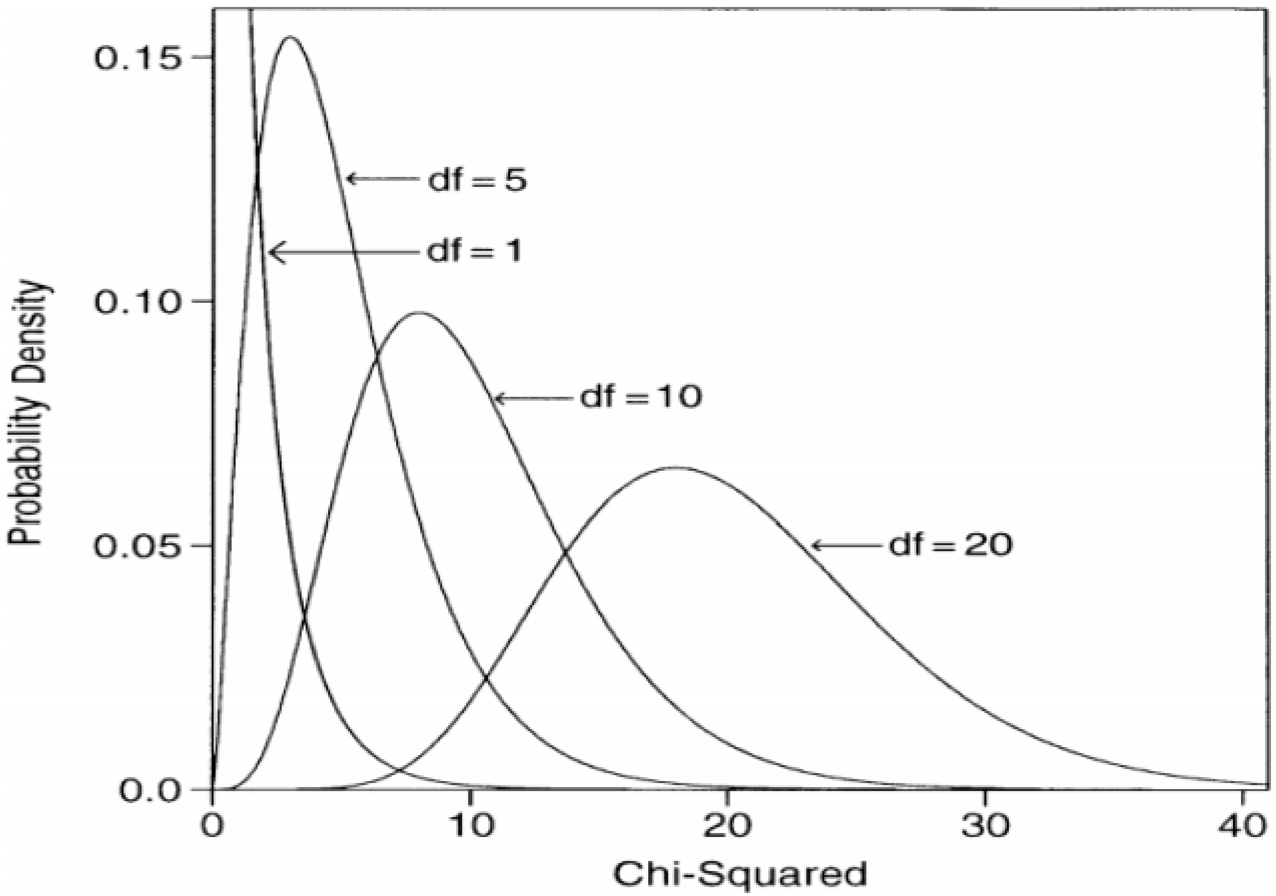
Degrees of Freedom (df) Significance Level (α)	0.01	0.05	0.10	0.25	0.50
1	6.635	3.841	2.706	1.323	0.454
2	9.210	5.991	4.605	2.773	1.386
3	11.345	7.815	6.251	3.930	2.366
4	13.277	9.488	7.779	5.178	3.357
5	15.086	11.070	9.236	6.571	4.351
6	16.812	12.592	10.645	7.962	5.348
7	18.475	14.067	12.017	9.364	6.346
8	20.090	15.507	13.362	10.773	7.344
9	21.666	16.919	14.684	12.189	8.343
10	23.209	18.307	15.987	13.603	9.342

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

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

$$E_i = \frac{Row\ Total \times Column\ Total}{Grand\ Total}$$

Chi-Square test χ^2



- The degrees of freedom for tests of hypothesis that involve an rxc contingency table is **equal to (r-1)x(c-1);**

Chi-Square test χ^2

Application of chi square test

1. **Goodness-of-fit:** uses frequency data from a sample to test hypotheses about the shape or proportions of a ***<u>population</u>***.
2. **Test for independence**:
    1. (2×2 chi-square test): Testing hypotheses about the relationship between two variables in a population,
    2. (a x b chi-square test) or (r x c chi-square test)

Chi-Square test χ^2

Q1. Given Eye colour in a sample of 40 people: Blue 12, brown 21, green 3, others 4
Given Eye colour in population: Brown 80%, Blue 10%, Green 2%, Others 8%
Is there any difference between proportion of sample to that of population (use alpha= 0.05)

Solution: Assume Sample is randomly selected from the population.
Null hypothesis: there is no significant difference in proportion of eye colour of sample to that of the population.
Alternative hypothesis: there is significant difference in proportion of eye colour of sample to that of the population.

Color	Sample frequency	Expected frequency
Blue	12	40*10/100= 4
Brown	21	40*80/100=32
Green	3	40*2/100=0.8
Others	4	40*8/100=3

$$\chi^2 = \frac{(12-4)^2}{4} + \frac{(21-32)^2}{32} + \frac{(3-0.8)^2}{0.8} + \frac{(4-3)^2}{3}$$

=(64/4) + (121/32)+(4.8/0.8)+(1/3)
=16+3.78+6+0.3
=26.08

Chi-Square test χ^2

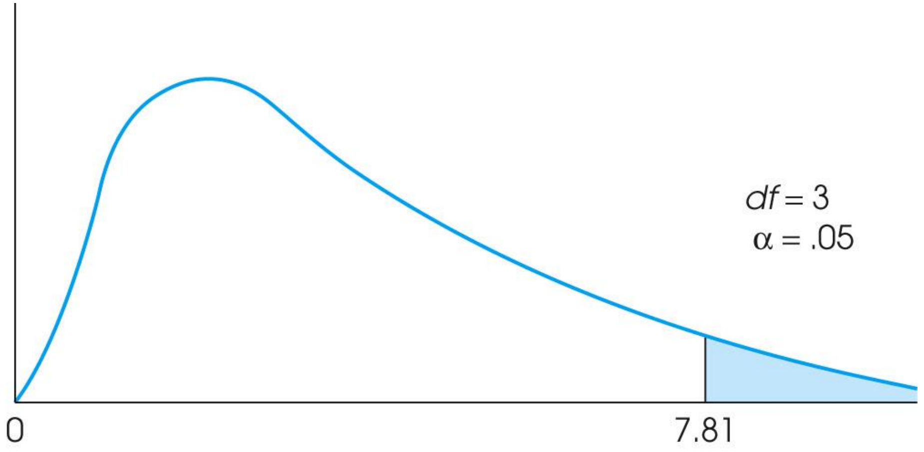
Q1. Given Eye colour in a sample of 40 people: Blue 12, brown 21, green 3, others 4
Given Eye colour in population: Brown 80%, Blue 10%, Green 2%, Others 8%
Is there any difference between proportion of sample to that of population (use alpha= 0.05)

Solution: Assume Sample is randomly selected from the population.

<u>Null hypothesis:</u> there is no significant difference in proportion of eye colour of sample to that of the population.

<u>Alternative hypothesis:</u> there is significant difference in proportion of eye colour of sample to that of the population.

α =0.05
d.f.(degree of freedom)=K-1 = 4-1 =3
(K=Number of subgroups)
critical value for α =0.5 and df=3 => 7.81



Chi-Square test χ^2

Calculated chi> tabulated chi

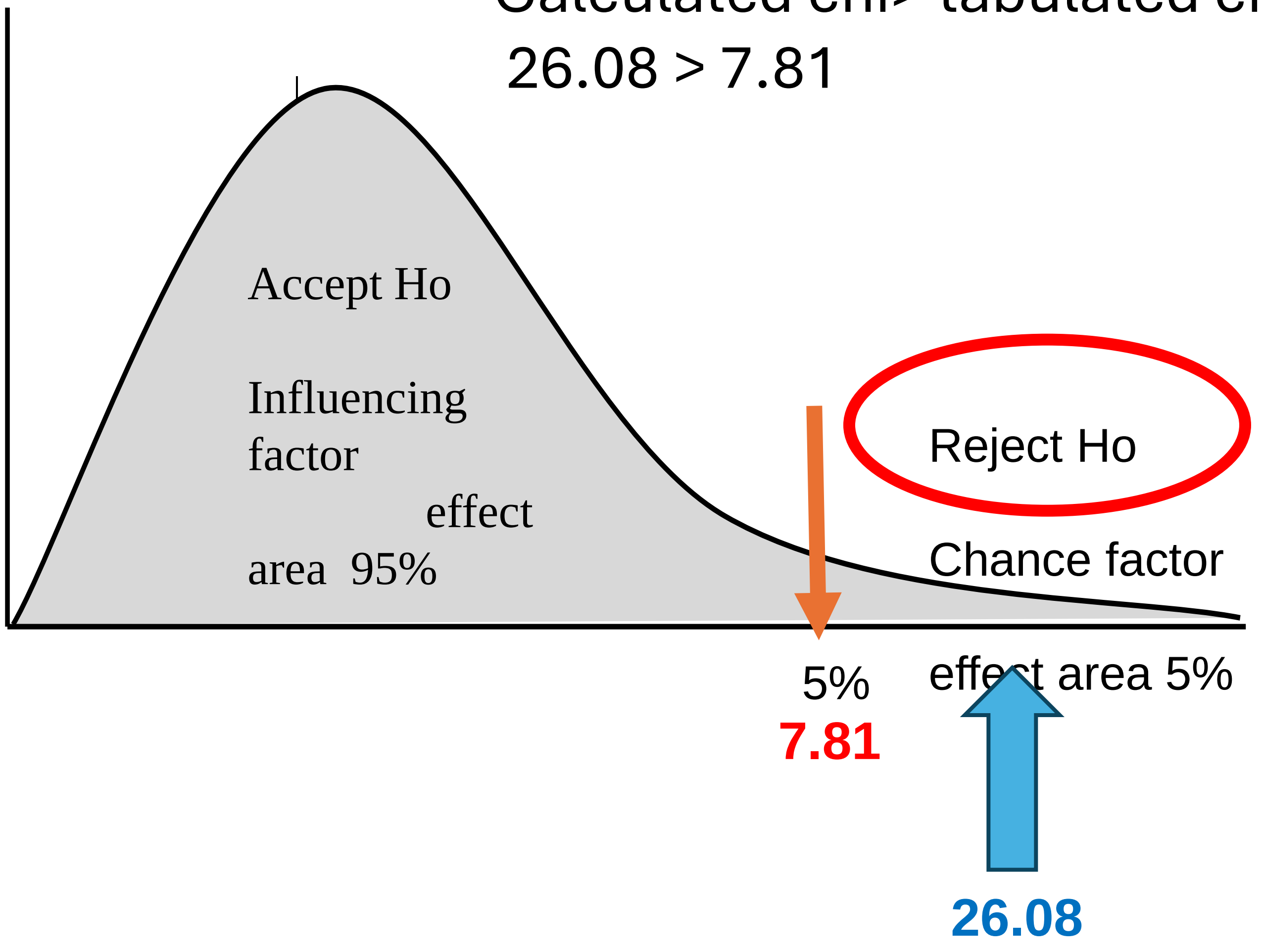
26.08 > 7.81

Conclusion: We reject H0 &accept HA

There is significant difference in

<u>**proportion**</u> of eye colour of sample to

that of the population.



Chi-Square test χ^2

Q2. A total 1500 workers on 2 operators (A&B) were classified as deaf & non-deaf according to the following table. Is there association (dependence) between deafness & type of operator. Let α 0.05

HO: there is no significant **<u>association</u>** between type of operator & deafness.
HA:there is significant **<u>association</u>** between type of operator & deafness.

Operator	deaf	Not deaf.	total
A	100	900	1000
B	60	440	500
total	160	1340	1500

Total number of items=1500

Total number of defective items=160

α =0.05
d.f.(degree of freedom)=(2-1)(2-1) = 1
critical value for α =0.5 and df=1 => 3.841

Chi-Square test χ^2

Q2. A total 1500 workers on 2 operators (A&B) were classified as deaf & non-deaf according to the following table. Is there association (dependence) between deafness & type of operator. Let α 0.05

Operator	deaf	Not deaf.	total
A	100	900	1000
B	60	440	500
total	160	1340	1500

Total number of items=1500

Total number of defective items=160

Expected deaf from Operator A = 1000 * 160/1500 = 106.7

(expected not deaf=1000-106.7=893.3)

Expected deaf from Operator B = 500 * 160/1500 = 53.3

$$\chi^2 = \frac{(100-106.7)^2}{106.7} + \frac{(900-893.3)^2}{893.3} + \frac{(60-53.3)^2}{53.3} + \frac{(440-446.7)^2}{446.7}$$

= 0.42+0.05+o.84+0.10 = 1.41

Chi-Square test χ^2

Q2. A total 1500 workers on 2 operators (A&B) were classified as deaf & non-deaf according to the following table. Is there association (dependence) between deafness & type of operator. Let α 0.05

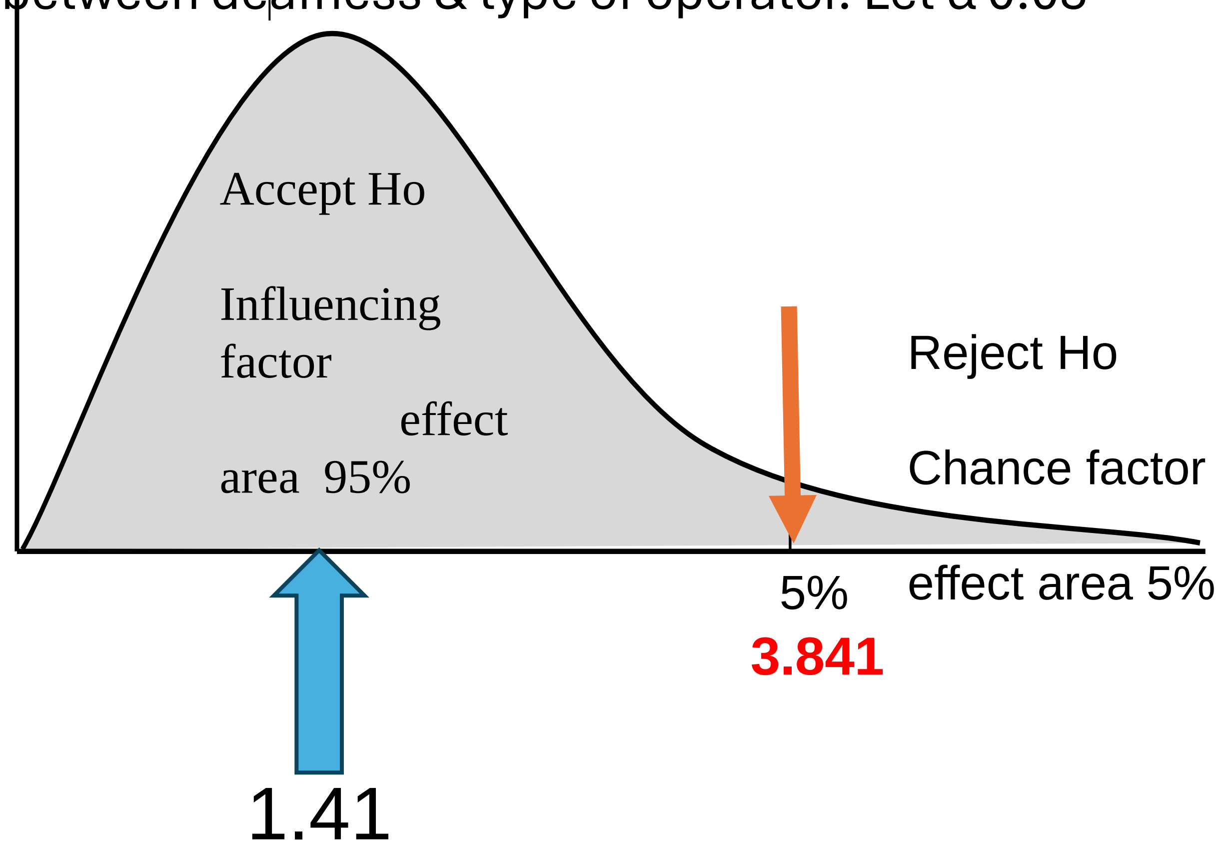
Calculated chi< tabulated chi

1.41 < 3.841

Conclusion: We accept H0

HO may be true

There is no significant **<u>association</u>** between type of operator & deafness.



Chi-Square test

Test for Independence using
(a x b chi-square test) or
(r x c chi-square test)

Calculation of expected frequencies: For r × c contingency table, the expected frequencies are as follow:

$$e_i = \frac{Row\ total(rt_i) \times Column\ total(ct_i)}{Grand\ total(n)}$$

Where e_i= expected frequency of cells and is $e_1, e_2, \ldots, e_k$ where k is the number of cells in the body of the table.

Consider the following 3 by 2 contingency table

Classification criteria 2	*Classification criteria 1*		
	Class 1	*Class 2*	*Total*
Category 1	a	b	$a+b$
Category 2	c	d	$c+d$
Category 3	e	f	$e+f$
Total	$a+c+e$	$b+d+f$	n

The expected value for the first cell (a), $e_1 = \frac{(a+b)(a+c+e)}{n}$

The expected value for the first cell (b), $e_2 = \frac{(a+b)(b+d+f)}{n}$

……………………………………………………………………………:

The expected value for the first cell (f), $e_6 = \frac{(e+f)(b+d+f)}{n}$

Chi-Square test χ^2

Q3. Perform a Chi-Square test to analyze the relationship between alcohol consumption (number of beers per day) and liver disease. The contingency table is given below.

Alcohol Drinking (No. of bottle beers/day)	Liver Disease Yes	Liver Disease No	Total
≤2	20	80	100
3-5	90	30	120
≥6	240	40	280
Total	350	150	500

1. State the Hypotheses:

•**Null Hypothesis (H0):** There is no association between the number of beers consumed per day and the presence of liver disease. The two variables are independent.

•**Alternative Hypothesis (H1):** There is an association between the number of beers consumed per day and the presence of liver disease. The two variables are not independent.

2. The expected frequency for each cell is calculated as:

(Row Total * Column Total) / Grand Total

Expected values are shown in brackets with each cell.

Beers/Day	Liver Disease (Yes)	Liver Disease (No)	Total
≤2	20 (70)	80 (30)	100
3-5	90 (84)	30 (36)	120
≥6	240 (196)	40 (84)	280
Total	350	150	500

3. Calculate the Chi-Square Statistic (χ^2):

$\chi^2 = \Sigma$ [(Observed - Expected)^2 / Expected]

= 35.71 + 83.33 + 0.43 + 1.00 + 2.21 + 7.74 = 153.4

4. Find critical value for df= (3-1)(2-1) = 2 and alpha = 0.05

Critical value = 5.991

5. Compare: Our calculated χ^2 (130.42) is much greater than the critical value (5.991). Therefore, we reject the null hypothesis.

Conclusion: There is a statistically significant association between the number of beers consumed per day and the presence of liver disease. The variables are not independent.