Chi-square

DR BISOLA ADEBAYO MPHFEP FMCPH

DEPARTMENT OF COMMUNITY HEALTH AND PHC

Learning Objectives

Perform the chi-square test manually

Interprete the results of chi-sqare tests

Introduction

The Chi-square test is one of the key tests conducted for hypothesis testing

A specific statement or hypothesis is made about a population parameter, and sample statistics are used to assess the likelihood that the hypothesis is true

The hypothesis is based both on available information and the investigator’s opinion about the population parameter

Background

- Was developed by Karl Pearson in 1900
- Chi-square is appropriate when the outcome is discrete (categorical, ordinal, dichotomous)
- The test follows a specific distribution known as the chi-square probability distribution
- In general, used to measure the difference between what is observed and what is expected according to an assumed hypothesis

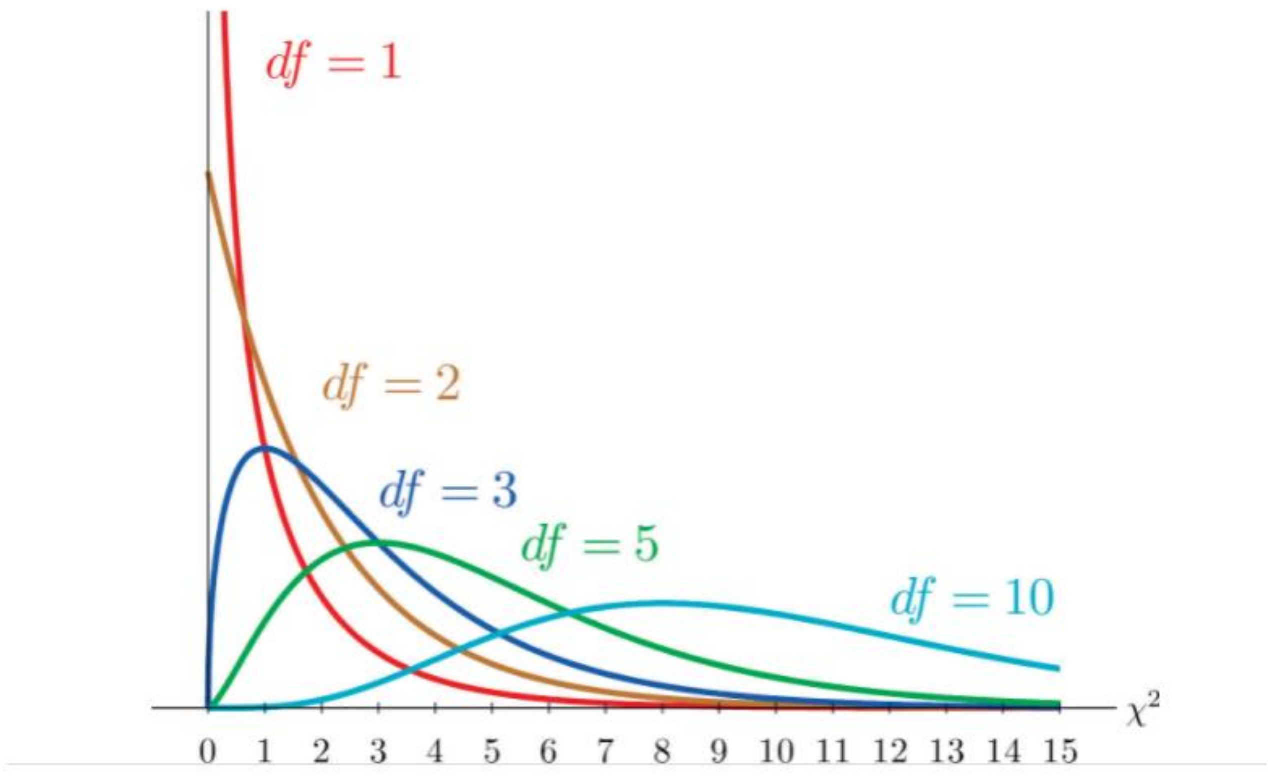
Characteristics of the Chi-Square test

- Non-parametric test as no rigid assumptions are necessary in regard to the type of population,
- Based on frequencies and not on the parameters like mean and standard deviation
- Used for testing the hypothesis and is not useful for estimation
- Commonly used in research

Chi-Square Distribution

- If X1, X2,….Xn are independent normal variates and each is distributed normally with mean zero and standard deviation unity, then X12+X22+……+Xn2= ∑ Xi2 is distributed as chi square (c2)with n degrees of freedom (d.f.) where n is large
- If degree of freedom > 2 : Distribution is bell shaped
- If degree of freedom = 2 : Distribution is L shaped with maximum ordinate at zero
- If degree of freedom <2 (>0) : Distribution L shaped with infinite ordinate at the origin.dd

Figure 11.1 Many χ2 Distributions



Chi-square Distribution

- The χ2 distribution is an asymmetric distribution that has a minimum value of 0, but no maximum value
- The curve reaches a peak to the right of 0, and then gradually declines in height, the larger the χ2 value is
- The curve approaches, but never quite touches, the horizontal axis
- For each degree of freedom there is a different χ2 distribution

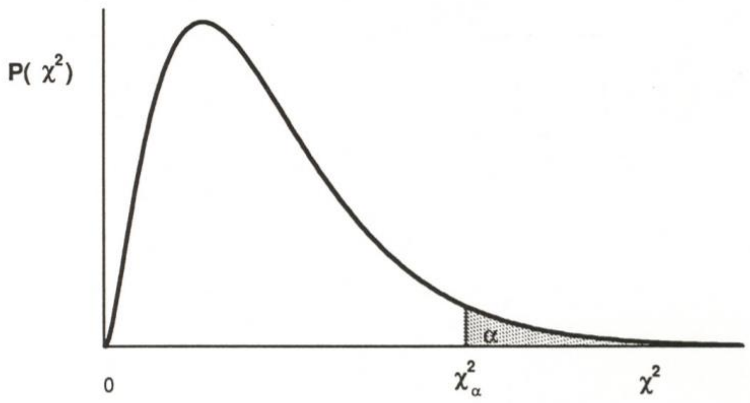


Figure J.1: The χ^2 distribution

Assumptions

At least 80% of the expected values in the distribution should be greater than 5

None of the expected values in the distribution should be less than 1

Observations are independent

Decision Rule

If the observed value of X^2 (x^2*obs*) is greater than or equal to the critical value of X^2(then we can reject H_0

Table used: Chi square distribution

Uses of Chi-square Test

- To describe the distribution of a sum of squared random variables
- To test the goodness of fit of a distribution of data,
- To test whether data series are independent
- For estimating confidences surrounding variance and standard deviation for a random variable from a normal distribution
- Test of homogeneity

Test of Independence of Attributes

- Test enables us to explain whether or not two attributes are associated or independent of each other

Conditions for application of the Chi-Square Test

The following conditions should be satisfied before X2 test can be applied

- The data must be in the form of frequencies
- The frequency data must have a precise numerical value and must be organized into categories or groups
- Observations recorded and used are collected on a random basis
- All the items in the sample must be independent
- No group should contain very few items
  - In case where the frequencies are less than 10, regrouping is done by combining the frequencies of adjoining groups so that the new frequencies become greater than 10. (Some statisticians take this number as 5, but 10 is regarded as better by most of the statisticians.)
- The overall number of items must also be reasonably large. It should normally be at least 50.

Calculating the test statistic

- The test statistic is:

$$c^2 = \sum_{i=1}^{k} \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

- The degrees of freedom are:
  - $(r-1)(c-1)$
  - r = # of rows and c = # of columns
- Where:
  - O_i = the observed frequency in the i^{th} cell of the table
  - E_i = the expected frequency in the i^{th} cell of the table

Steps in calculating Chi-square

1. State the null and alternative hypotheses
2. State the decision rule
3. Calculate expected frequency for all cells in the chi-square table using
4. Calculate the chi-square value (obtained chi-square), using the observed (Oi) and expected values (Ei)
5. Find the critical chi-square value
6. Apply the decision rule
7. State the practical conclusion

Steps in calculating Chi-square

Note the following

1. The formula for calculating expected values is

$$= \frac{row\ total \times column\ total}{population\ total}$$

2. Formula for Chi-square $X^2 = \sum_{i=1}^{k} \left[\frac{(O_i - E_i)^2}{E_i} \right]$

3. Finding the critical chi-square value first requires calculating the degrees of freedom (df)

df = (# rows – 1) (# columns – 1)

Guideline for interpreting the Chi-Square test

- The χ^2 statistic is calculated under the assumption of no association

- **Large value of χ^2 statistic** $\Rightarrow$ small probability of occurring by chance alone ($p < 0.05$) $\Rightarrow$ conclude that **association** exists between disease and exposure

- **Small value of χ^2 statistic** $\Rightarrow$ large probability of occurring by chance alone ($p > 0.05$) $\Rightarrow$ conclude that **no association** exists between disease and exposure

Limitations of a Chi-Square Test

- The data is from a random sample
- This test will not give a reliable result with one degree of freedom if the expected value in any cell is less than 10 (taken as 5 by some statisticians)
- In such case, Yate's correction is necessary
- In contingency tables larger than 2 by 2, Yate's correction cannot be applied
- Interpret this test with caution if sample total or total of values in all the cells is less than 50
- The test tells the presence or absence of an association between the events but doesn't measure the strength of association
- This test doesn't indicate the cause and effect, it only tells the probability of occurrence of association by chance
- The test is to be applied only when the individual observations of sample are independent which means that the occurrence of one individual observation (event) has no effect upon the occurrence of any other observation (event) in the sample under consideration

Practice question

- A researcher is interested in understanding whether there is an association between smoking status (smoker vs. non-smoker) and the development of respiratory infections (yes vs. no) among patients in a clinic.
- The researcher collects data from 200 patients and obtains the following contingency table:

	Respiratory Infection	No Respiratory Infection
Smoker	30	70
Non-Smoker	20	80

Solution

State the Null Hypothesis

Null Hypothesis (H_0):

- There is no association between smoking status and the development of respiratory infections.
- Development of respiratory infections is independent of smoking status

State the alternate hypothesis

Alternative Hypothesis (H_1):

- There is an association between smoking status and the development of respiratory infections.
- Development of respiratory infections is independent of smoking status

Solution

- State your decision rule based on your alpha and your degree of freedom
- Degree of freedom = (r -1)(c-1)

  (2-1)(2-1) = 1 x 1 = 1

- Alpha value (α) is usually given = 0.05
- Decision rule = With an α of 0.05, If the calculated chi-square value is greater than the critical value, reject the null hypothesis
- If $\chi^2_{calc} \geq \chi^2_{crit}$, reject H_0
- If $\chi^2_{calc} < \chi^2_{crit}$, fail to reject H_0

Complete your contingency table

	Respiratory Infection	No respiratory infection	Total
Smoker	30	70	100
Non-smoker	20	80	100
Total	100	100	200

Calculate the expected frequencies

Observed (O)	Expected (E)	Expected (E)	O - E	$(O - E)^2$	$(O - E)^2/E$	$(O - E)^2/E$
30	(100x50)/200	25	30 – 25 = 5	25	25/25	1.000
70	(100x150)/200	75	70 – 75 = -5	25	25/75	0.333
20	(100x50)/200	25	20 – 25 = -5	25	25/25	1.000
80	(100x150)/200	75	80 – 75 = 5	25	25/75	0.333
					$\sum(O_i - E_i)^2 / E_i = 2.666$	

Chi-square table

Degrees of freedom (*df*)	Significance level (α)							
	.99	.975	.95	.9	.1	.05	.025	.01
1	-------	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000

Determine the critical value

$\chi^2_{calc} = 2.666$

With df = 1, $\alpha = 0.05$

$\chi^2_{crit} = 3.841$

Apply the Decision rule

Compare χ2 to the Critical Value:

$\chi^2_{calc} = 2.666$

$\chi^2_{crit} = 3.841$

Since 2.666 < 3.841, we **fail to reject the null hypothesis**

State the practical conclusion

There is **no statistically significant association** between smoking status and the development of respiratory infections at the 0.05 significance level

This means the observed differences in the data could be due to random variation (chance), and we do not have sufficient evidence to conclude that smoking status affects the likelihood of respiratory infections