Chi-square Test and its Application in Hypothesis Testing

Rakesh Rana, Richa Singhal

Statistical Section, Central Council for Research in Ayurvedic Sciences, Ministry of AYUSH, GOI, New Delhi, India

Abstract

In medical research, there are studies which often collect data on categorical variables that can be summarized as a series of counts. These counts are commonly arranged in a tabular format known as a contingency table. The chi-square test statistic can be used to evaluate whether there is an association between the rows and columns in a contingency table. More specifically, this statistic can be used to determine whether there is any difference between the study groups in the proportions of the risk factor of interest. Chi-square test and the logic of hypothesis testing were developed by Karl Pearson. This article describes in detail what is a chi-square test, on which type of data it is used, the assumptions associated with its application, how to manually calculate it and how to make use of an online calculator for calculating the Chi-square statistics and its associated *P*-value.

Key words: Categorical data analysis, Chi-square test, hypothesis testing, online calculator

The logic of hypothesis testing was first invented by Karl Pearson (1857–1936), a renaissance scientist, in Victorian London in 1900.[1] Pearson's Chi-square distribution and the Chi-square test also known as test for goodness-of-fit and test of independence are his most important contribution to the modern theory of statistics. The importance of Pearson's Chi-square distribution was that, the statisticians could use the statistical methods that did not depend on the normal distribution to interpret the findings. He invented the Chi-square distribution to mainly cater the needs of biologists, economists, and psychologists. His paper in 1900 published in Philosophical magazine elaborates the invention of Chi-square distribution and goodness of fit test.[2,3]

Chi-square test is a nonparametric test used for two specific purpose: (a) To test the hypothesis of no association between two or more groups, population or criteria (i.e. to check independence between two variables); (b) and to test how likely the observed distribution of data fits with the distribution that is expected (i.e., to test the goodness-of-fit). It is used to analyze categorical data (e.g. male or female patients, smokers and non-smokers, etc.), it is not meant to analyze parametric or continuous data (e.g., height measured in centimeters or weight measured in kg, etc.).

For example if we want to test that in a health camp attended by 50 persons the one who exercise regularly are having lesser body mass index (BMI) by taking their actual BMI values, than it cannot be tested using a Chi-square test. However, if we divide the same set of 50 persons into two categories as obese with BMI ≥ 30 and nonobese with BMI < 30, than the same data can be tested using a Chi-square test by counting the number of obese and nonobese persons across two groups, the one who exercise regularly and the one who does not. A 2x2 contingency table also known as cross tables can be constructed for calculating a Chi-square statistic [Table 1].

Access this article online

Quick Response Code:

Website: www.j-pcs.org

DOI: 10.4103/2395-5414.157577

Assumptions Underlying a Chi-square Test

- The data are randomly drawn from a population
- The values in the cells are considered adequate when expected counts are not <5 and there are no cells with zero count[4,5]
- The sample size is sufficiently large. The application of the Chi-square test to a smaller sample could lead to type II error (i.e. accepting the null hypothesis when it is actually false). There is no expected cut-off for the sample size; however, the minimum sample size varies from 20 to 50
- The variables under consideration must be mutually exclusive. It means that each variable must only be counted once in a particular category and should not be allowed to appear in other category. In other, words no item shall be counted twice.

How to Calculate a Chi-square Statistics?

The formula for calculating a Chi-square statistic is:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Address for correspondence: Dr. Richa Singhal, Central Council for Research in Ayurvedic Sciences, Ministry of AYUSH, GOI, New Delhi, India.
E-Mail: richa.singhal2k@gmail.com

Where,
O stands for the observed frequency,
E stands for the expected frequency.

Expected count is subtracted from the observed count to find the difference between the two. Then the square of the difference is calculated to get rid of the negative vales (as the squares of 2 and −2 are, of course, both 4). Then the square of the difference is divided by the expected count to normalize bigger and smaller values (because we don't want to get bigger Chi-square values just because we are working on large data sets). The sigma sign in front of them denotes that we have, to sum up, these values calculated for each cell.

As an example, suppose we want to find out that whether there is an association between smoking and lung disease.

The null and alternative hypothesis will be:

H_0: There is no association between smoking and lung disease.

H_1: There is an association between smoking and lung disease.

The basic step for calculating a Chi-square test is setting up a 2 × 2 contingency table [Table 2].

The general formula for calculating the expected counts from observed count for a particular cell is [(corresponding row total * corresponding column total) /Total no. of patients] [Table 3].

Before we proceed further, we need to know how many degrees of freedom (df) we have. When a comparison is made between one sample and another, a simple rule is that the df equals (number of columns − 1) × (number of rows − 1) excluding the rows and column containing the total. Hence, in our example df = (2−1) × (2−1) = 1.

Hypothetical data for calculating the Chi-square test for our example of testing an association between smoking and lung disease is given in Table 4. Chi-square test can be calculated manually by using the formula described above. Refer Tables 5 and 6 for manual calculations. Chi-square value for our example as shown in Table 6 is 3.42, df = 1. If we want to test our hypothesis at 5% level of significance than our predetermined alpha level of significance is 0.05. Looking into the Chi-square distribution table [Table 7] with 1 degree of freedom and reading along the row we find our value of χ^2 (3.42) lies between 2.706 and 3.841. The corresponding probability is between the 0.10 and 0.05 probability levels. That means that the *P* value is above 0.05 (it is actually 0.065). Since a *P* value of 0.065 is greater than the conventionally accepted significance level of 0.05 (i.e., $P > 0.05$) we fail to reject the null hypothesis or in other words we accept our null hypothesis and conclude that there is no association between smoking and lung disease.

How to Use a Chi-square Distribution Table to Approximate P Value?

Scientists and statisticians use large tables of values to

Table 1: Sample of a 2×2 contingency table

	Obese	Nonobese	Total
Exercise regularly	7	20	27
Don't exercise regularly	15	8	23
Total	22	28	50

Table 2: General notation for a 2×2 contingency table (observed values for the data)

	Smokers	Nonsmokers	Total
Suffering from lung disease	a	b	a+b
Not suffering from lung disease	c	d	c+d
Total	a+c	b+d	a+b+c+d=*n*

Table 3: Expected values for the data presented in a 2×2 table

	Smokers	Nonsmokers
Suffering from lung disease	(a+b) (a+c)/*n*	(a+b) (b+d)/*n*
Not suffering from lung disease	(a+c) (c+d)/*n*	(c+d) (b+d)/*n*

Table 4: Hypothetical data containing observed values for calculating Chi-square statistics

	Smokers	Nonsmokers	Total
Suffering from lung disease	36	14	50
Not suffering from lung disease	30	25	55
Total	66	39	105

Table 5: Expected values for the hypothetical data

	Smokers	Nonsmokers
Suffering from lung disease	66×50/105=31.42	39×50/105=18.57
Not suffering from lung disease	66×55/105=34.57	39×55/105=20.43

Table 6: Summarizing the data for calculating the Chi-square value

Observed count (O_i)	Expected count (E_i)	(O_i-E_i)	(O_i-E_i)^2	(O_i-E_i)^2/E_i
36	31.42	4.57	20.90	0.66
14	18.57	−4.57	20.90	1.13
30	34.57	−4.57	20.90	0.60
25	20.43	4.57	20.90	1.02
χ^2				3.42

Table 7: Excerpts from the Chi-square distribution table

	Probability level (alpha)					
df	0.5	0.10	0.05	0.02	0.01	0.001
1	0.455	2.706	3.841	5.412	6.635	10.827
2	1.386	4.605	5.991	7.824	9.210	13.815
3	2.366	6.251	7.815	9.837	11.345	16.268
4	3.357	7.779	9.488	11.668	13.277	18.465
5	4.351	9.236	11.070	13.388	15.086	20.517

calculate the *P* value for their experiment. These tables are generally set up with the vertical axis on the left corresponding to df and the horizontal axis on the top corresponding to *P* value. Use these tables by first finding our df, then reading that row across from the left to the right until we find the first value bigger than our Chi-square value. Look at the corresponding *P* value at the top of the column. Chi-square distribution tables are available from a variety of sources-they can easily be found online or in science and statistics textbooks.

Using an Online Chi-square Calculator

The Chi-square statistics and its associated *P* value can be calculated through online calculators also which are easily available on the internet. For user-friendly online calculator, you may visit this uniform resource locator www.socscistatistics.com/tests/chisquare/default2.aspx. Many more online calculators are available on the World Wide Web. The basic step for using an online calculator is to correctly fill in your data into it.

Step by step procedure of using an online calculator is described below:

- Step 1: For our example of finding an association between smoking and lung disease we have to fill in the observed values in the cells of an online calculator [Figure 1]
- Step 2: Click on the next button. Another screen will pop up as shown in Figure 2
- Step 3: Click on the Calculate Chi^2 button. And you are done with your calculation Output of the Chi-square test will be as shown in Figure 3.

The image above shows the Chi-square value as 3.4177 and its associated *P* value as 0.0645 which is actually greater than *P* value of 0.05, hence no significant difference has been observed. To conclude, there is no association between smoking and lung disease.

What Does a Chi-square Test Tell and What it Does Not?

It may be clearly understood that Chi-square test only tells us the probability of independence of a distribution of data or in simple terms it will only test that whether two variables are associated with each other or not. It will not tell us that how closely they are associated. For instance in the above example, the Chi-square test will only tell us that whether there is any relation between smoking and lung disease. It will not tell us that how likely it is, that smokers are prone to lung disease. However, once we got to know that there is a relation between these two variables, we can explore other methods to calculate the amount of association between them.

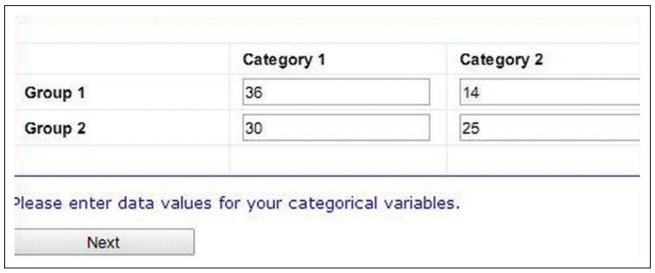

Figure 1: Setting up the data in the 2 × 2 table of an online calculator.

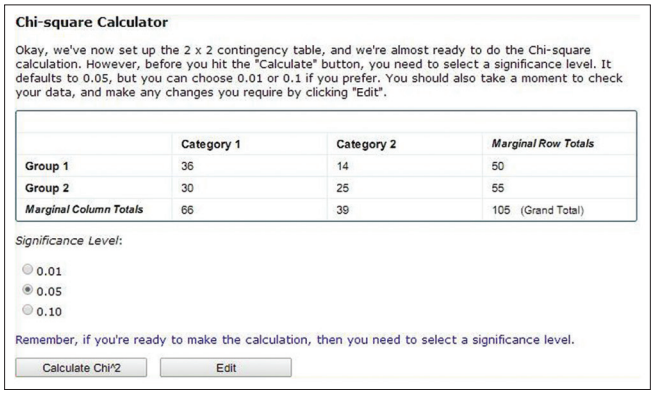

Figure 2: Setting up the significance level for calculation.

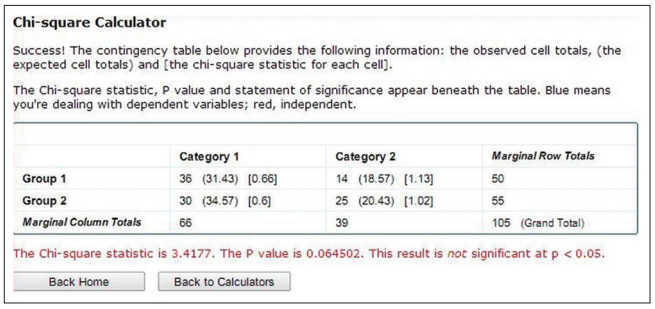

Figure 3: Chi-square value and *P*-value calculated by online calculator.

References

1. Magnello ME Karl Pearson and the origin of modern statistics: An elastician becomes a statistician, Rutherford J, Vol. 1, 2005-2006. Available online at: http://rutherfordjournal.org.
2. Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philos Mag Ser 1900;50:157-75.
3. Plackett RL. Karl Pearson and the Chi-squared test. Int Stat Rev 1983;51:59-72.
4. Yates F, Moore D, McCabe G. The Practice of Statistics 1st ed. New York: W.H.Freeman, 1999.
5. Yates F. Contingency table involving small numbers and the Chi-squared test. Suppl J R Stat Soc 1934;1:217-35.

How to cite this article: Rana R, Singhal R. Chi-square test and its application in hypothesis testing. J Pract Cardiovasc Sci 2015;1:69-71.

Source of Support: Nil. **Conflict of Interest:** None declared.