CHI-Squared Test of Independence

Minhaz Fahim Zibran
Department of Computer Science
University of Calgary, Alberta, Canada.

Email: mfzibran@ucalgary.ca

Abstract

Chi-square (X^2) test is a nonparametric statistical analyzing method often used in experimental work where the data consist in frequencies or 'counts' – for example the number of boys and girls in a class having their tonsils out – as distinct from quantitative data obtained from measurement of continuous variables such as temperature, height, and so on. The most common use of the test is to assess the probability of association or independence of facts [3]. This paper summarizes the chi-squared test. Beginning from the basics it discusses with descriptive examples on where and how to apply this test. The purpose of the paper is to present a quick overview on chi-square test, so that one who doesn't have much knowledge on statistics may use it as a beginner's guide.

Keywords: chi-square, statistical test, qualitative data, test of independence, test of association

1 Introduction

Suppose we have a sample of boys and girls from the 5th, 8th, and 12th grade of school. We may want to know whether there is an association between the gender of students and the grade levels. Or, we may want to know whether the men and women in a liberal arts college differ in their section of majors. These are the type of questions that the chi-squared (X^2) test is designed to answer. It was first introduced by British statistician Karl Pearson in 1900 [1, 2].

Before going into details of chi-squared test section 2 first presents some statistical preliminaries necessary to clearly understand it. Section 3 then illustrates the chi-square test in detail with a simple example. Section 4 points out some limitations of chi-square test. Finally I conclude the paper in section 5, where I also provide direction to further reading. The appendix includes an abridged form of the standard chi-square table.

2 Preliminaries

Population: A population is an individual or group that represents all the members of a certain group or category of interest [4]. Populations do not have to include people. Suppose we want to know the average age of the cats in the city. The population in this study is made up of cats, not people. A population does not need to be large to count as a population. We may want to know the average height of the 3 kids in a family. In this study, the population is comprised of only 3 kids.

Sample: A sample is a subset of a given population. Samples are not necessarily good representations of the populations from which they are selected. But choosing representative samples is important for most experiments.

Parameter: A parameter is a value generated from, or applied to a population.

Variable: A variable is pretty much anything that can be codified and can have value from a set (domain) of more than one values. Variables may be quantitative or qualitative. A *quantitative variable* is one that is scored in such a way that its values indicate some sort of amount. For example, height is a quantitative variable. On the contrary, a *qualitative variable* indicates some kind of category. A commonly used qualitative variable in social science research is the *dichotomous variable*, which has two different categories. For instance, gender has two categories: male and female. Chi-square test is applicable to when we have qualitative variables classified into categories.

Nominally scaled variable: A nominally scaled variable is one in which the labels that are used to identify the different levels of the variable have no weight, or numeric value [4]. For example the sample may be divided into two groups labeled "0" and "1". In this case the value "1" does not indicate a higher score than the value of "0". Rather the, "0" and "1" are simply names, or labels have been assigned to each group.

Contingency table: When the members of a sample are doubly classified (i.e., classified in two separate ways), the results may be arranged in rectangular tables. Such a table is called *contingency table.* For example, in 1956, the number of people on record who died of tuberculosis in England and Wales was 5375. Of these 3804 were males and 1571 were females: 3534 males and 1319 females died of tuberculosis of respiratory system, while the remainder died of other forms of tuberculosis. There data can be arranged in contingency table as shown in Table 1[1]. This 2 × 2 (the members of the sample having been dichotomized in two different ways) contingency table is an example of the simplest form.

[1] this example and table is actually taken from [3]

	Males	Females	Total
Tuberculosis of respiratory system	3534	1319	4853
Other forms of tuberculosis	270	252	522
Tuberculosis (all forms)	3804	1571	5375

Table 1: Observed frequencies of deaths from Tuberculosis in England and Wales in 1956

	Males	Females	Total
Tuberculosis of respiratory system	E_1	E_2	4853
Other forms of tuberculosis	E_3	E_4	522
Tuberculosis (all forms)	3804	1571	5375

Table 2: Table 1 with cells replaced by letters

3 Chi-squared Test

The entries in the *cells* in a contingency table may be frequencies, as in Table 1, or frequencies may be transformed into proportions or percentages. However, it is important to note that in whatever form (frequencies, proportions, etc.) they are presented are not continuous measurements. Chi-squared test can be applied to only discrete data [3]: for the purpose of the test, of course, continuous data can be often put into discrete form by the use of intervals on a continuous scale. For instance, age is a continuous variable, but if people are classified into different age-groups, then the intervals of time corresponding to these groups can be treated as if they were discrete units.

Chi-square (X^2) test is a nonparametric statistical test to determine if the two or more classifications of the samples are independent or not. For explanation, let us consider the data presented in Table 1 where the people are classified according to two attributes: gender and type of tuberculosis. We may want to determine whether death caused by tuberculosis of respiratory system or other type of tuberculosis is dependent on gender. To get the answer, we may apply chi-square test. A. E. Maxwell [3] presented this example with the data shown in Table 1 to elaborately describe the procedure of chi-square test. In this paper I use the same example and illustrate the procedure in a concise form.

In order to carry out the chi-square test, it will be helpful to rearrange Table 1 leaving the marginal frequencies as they are but replacing the frequencies in the cells of the body of the table by the letters E_1 to E_4. After this arrangement we get Table 2.

Looking at the column of the marginal totals on the right of the table we see that the proportion of deaths, males and females combined, due to tuberculosis of the respiratory system, is

$$\frac{4853}{5375} = 0.903 \tag{1}$$

	Males	Females	Total
Tuberculosis of respiratory system	3434.6	1418.4	4853
Other forms of tuberculosis	369.4	152.6	522
Tuberculosis (all forms)	3804	1571	5375

Table 3: Expected frequencies on the assumption of independent classification

Now, if the two classifications are independent, that is if the form of tuberculosis from which people die is independent of gender, we would expect that the proportion of males that died from tuberculosis of respiratory system would be equal to that of females died from the same cause, and consequently would equal the proportion of the total group, 0.903.

The expected values E_1 and E_2 then must be chosen so that the following holds.

$$\frac{E_1}{3804} = \frac{E_2}{1571} = \frac{4853}{5375} = 0.903 \tag{2}$$

Using equation 2 we find

$$E_1 = \frac{(4853 \times 3804)}{5375} = 3434.6 \tag{3}$$

Once the value of E_1 is known, the values of E_2, E_3 and E_4 can be deduced since the following facts are true.

$$E_1 + E_2 = 4853 \tag{4}$$

$$E_3 + E_4 = 522 \tag{5}$$

$$E_1 + E_3 = 3804 \tag{6}$$

$$E_2 + E_4 = 1571 \tag{7}$$

Calculating values of E_1, E_2, E_3, and E_4 and putting these expected values replacing the corresponding letters in Table 2, we get Table 3. These are the values that one would expect to find in the cells in the body of Table 1 were the two methods of classification independent. Though the number of people may not be fractional, fractional values may appear in the table of expected frequencies. Specially when the sample size is small, we retain the fractional values to increase the accuracy of subsequent calculations.

Now, if we refer again to the data, we see that the observed frequencies in Table 1 differs considerably from the expected frequencies in Table 3. The question in concern is whether this difference is such as could have arisen from random sampling error alone, or whether it indicates a real difference between genders: males and females.

Now, let us define our null hypothesis as follows.

Null hypothesis: The number of men and women died in 1956 due to tuberculosis of respiratory system and other types of tuberculosis is independent of their sex.

O	E	$(O-E)$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
3534	3434.6	99.4	9880.36	2.88
1319	1418.4	-99.4	9880.36	6.97
270	369.4	-99.4	9880.36	26.75
252	152.6	99.4	9880.36	64.75
5375	5375.0	0.0		$X^2 = 101.35$

Table 4: Calculating chi-square (X^2) for data in Table 1 and Table 3

Chi-square test, if properly applied may give us the answer by rejecting the null hypothesis or failing to reject it. The test is based on the chi-square (X^2) distribution. To compare the observed and expected frequencies, we produce chi-square (X^2) value using the formula stated in equation 8.

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{8}$$

In this equation 8, O_i stands for observed frequencies, E_i stands for expected frequencies, and i runs from $1, 2, \ldots, n$, where n is the number of cells in the contingency table.

To perform the calculations it is useful to arrange the observed and expected frequencies as shown in Table 4.

To asses the significance of the calculated value of X^2, we refer to the standard chi-square table presented in Appendix. This table contains the critical X^2 values on different *degrees of freedom* and levels of probability. Referring back to Table 3, we recall that once the value of any one of the E_i ($i = 1, \ldots, 4$) had been determined, all other E_i's could be deduced. In other words, when the marginal totals of a 2×2 contingency table is given, only one cell in the body of the table can be filled arbitrarily. This fact is expressed by saying that a 2×2 contingency table has only one degree of freedom. The degree of freedom (df) of a contingency table with r rows and c columns is computed using the following formula given in equation 9.

$$df = (r-1)(c-1) \tag{9}$$

To assess the significance of our chi-square value $X^2 = 101.35$, we enter the chi-square table of the Appendix, with $df = 1$, that is we look into the first row, which corresponds to one degree of freedom. The largest value in that row is 10.828 under the probability (P) level 0.001. A value of chi-square equal to or greater than 10.828 would be expected to occur by chance only once in a thousand times if the null hypothesis is true. Since our chi-square value 101.35 is much greater than 10.828, it would be expected to occur even less frequently. Hence our chi-square test rejects the null hypothesis. So, we conclude that the proportion of males died from tuberculosis of respiratory system, namely $\frac{3534}{3804} = 0.929$, is significantly different from the proportion of females, namely $\frac{1319}{1571} = 0.840$, that died from the same cause.

Before accomplishing the test we define a level of confidence, that is the probability level (P) we are going to accept. Once the X^2 value is computed and the number of degrees of freedom is determined, we go to the chi-square table and look into the row corresponding to the given degree of freedom. Then if we find our X^2 value to be less than (to the left side of our level of confidence) that of the value corresponding to our level of confidence, we conclude that our null hypothesis is probably true. On the contrary, if our X^2 value lies over the level of confidence or to its right indicating less probability of occurring the difference by chance, we know that our chi-square test rejects the null hypothesis. Therefore, we conclude that the classifications on population are dependent on each other.

4 Limitations of Chi-square Test

As mentioned before, chi-square test cannot be applied on continuous data. It can only be applied to qualitative data classified into categories, or labeled using nominally scaled variables [3, 4]. In the standard chi-square table presented in the Appendix the chi-square values computed using the formula in equation 8 assuming that the expected values (E) are large. Most statisticians warn against using the test when any of the expected values are less than 5. This warning implies that the use of chi-square test is restricted to large samples. However, where small samples are concerned, the difficulty can be overcome in a number of ways. One way is to apply a *correction of continuity*, known as *Yates correction*. Another way is using *Fisher's Exact Test*. Cochran [5] recommends that Fisher's exact test should be used when the total sample size in a 2×2 contingency table is less than 20, or when the sample is less than 40 and one of the expected frequencies is less than 5. More about these methods and chi-square test itself are nicely described in [3].

5 Concluding Remarks

Chi-square test tells us whether the classifications on a given population are dependent on each other or not. However, it is important to stress that the establishment of statistical association by means of chi-square does not necessarily imply any causal relationship between the attributes being compared, but it does indicate that the reason for the association is worth investigating [3]. For example, if further investigation carried out on the case of people's death from tuberculosis, we might have found out that the reason why the number of men died from tuberculosis of the respiratory system is higher than the women is the fact that there are more smokers in men than in women.

More on chi-square test may be found in the references given below. Maxwell in his book "Analysing Qualitative Data" [3] elaborately describes the chi-square test and related topics. The book by Timothy C. Urdan [4] covers necessary preliminary knowledge on statistics and illustrates the chi-square test with descriptive examples.

References

[1] J. C. W. Rayner and D. J. Best. *Smooth Tests of Goodness of Fit.* Oxford University Press, Inc., 1989. ISBN 0-19-505610-8.

[2] William Mendenhall, Robert J. Beaver, and Barbara M. Beaver. *Introduction to Probability and Statistics.* Brooks/Cole, a division of Thomson Learning, Inc., 2003. ISBN 0-534-39519-8.

[3] A. E. Maxwell. *Analysing Qualitative Data.* 4th Edition. Chapman and Hall Ltd., 1971. Library of Congress Catalog Card Number 75–10907.

[4] Timothy C. Urdan. *Statistics in Plain English.* 2nd Edition. Lawrence Erlbaum Associates, Inc., London, 2005.

[5] W. G. Cochran. *Some methods of strengthening the common x^2 tests.* Biometrics, 10, 417-51. 1954.

Appendix

Percentage points of chi-square distribution [2]

	Probability (P) level			
D.F	0.050	0.025	0.010	0.001
1	3.841	5.024	6.635	10.828
2	5.991	7.378	9.210	13.816
3	7.815	9.348	11.345	16.266
4	9.488	11.143	13.277	18.467
5	11.071	12.833	15.086	20.515
..	..	..	..	..
..	..	..	..	..

[2] Part of the chi-square table is presented here. The table with more values may be found in books on statistics. The full table may be found in Table 8 of *Biometrika Tables of Statisticians*, vol. 1, Biometrika Trustees.