CHI-SQUARE TESTS

mkhulekeli Nkosi 2017159092

2025-02-22

# The Chi – squared test.

This essay focuses on defining the Chi – square test, the assumptions that need to be met before applying the test and some misconceptions about the Chi – square test. It further applies the Chi – square test to simulated data which is assumed to be collected from a community to solve a cultural problem and meets the assumptions.

The Chi – square test is a non – parametric statistic, also called a distribution free test and this tests are used when one of the following pertains to the data, the level of measurement of all the variables is nominal (has two or more categories which have no order or ranking) or ordinal (categories have a meaningful order or ranking); sample sizes of the groups are unequal; if the original data were measured interval or ratio level then it must violate one of the assumptions of parametric test : normality(distribution of data is skewed), equal variance, and data is no longer interval or ratio(quantitative data measured at a continuous scale) (ML 2013). If one of the above conditions holds it means the data is non – parametric and non – parametric tests can be used.

If data is non – parametric, to apply the Chi – square test the assumptions listed below must be met (ML 2013).

1. Data in cells should be frequencies or counts of cases.  
  
2. The level or categories of variables are mutually exclusive (subject fits in only one level).  
  
3. Each subject may contribute data to one and only one cell in the chi – square.  
  
4. Study groups must be independent.  
  
5. Random selection – the data is obtained through random selection to ensure the sample is representative of the population with each cell with more than 5 observations.

The Chi – squared test is used to test for three things, **the goodness of fit**, **for independence** and **homogeneity of samples** and the same formula is used for each test to compute the test statistic is given as

, is the observed value, is the expected value number of observations. The obtained test statistic is compared against the critical value from the chi – square distribution with degree of freedom with as the number of columns and the number of rows for test of independence and for goodness of fit the degree of freedom is . Each test is distinct with specific hypothesis, interpretations and options following rejection of the null hypothesis (Franke, Ho, and Christie 2012).

# The Chi – square goodness of fit test.

This test is used when as sample is compared on a variable of interest against a population with known parameters with the null hypothesis as “The data follows a specific distribution” and the Alternative as “The data do not follow the specific distribution”(Franke, Ho, and Christie 2012). Many websites show how to calculate the test statistic by hand given a variable with low categories (Z. Bobbitt 2023 May 11). In this essay the function Chisq.test() is used in Rstudio.One of the examples looked at is where the assumption of independence and the expected value n each category are greater than 5 and also those of non-parametric are met and no.

**Example 1** aim is to see which group participates more in a community through attending community meetings.

set.seed(2017159092)  
n= 50  
#Y= 16 to 25 years  
#YA = 26 to 35 years of age  
#A = 36 - 45 years  
# O = 46 - 55 years  
# OO = 56 years and greater  
  
# H\_0 : we expect to be the same number of participation in each group.  
# H\_1 : at least one group differes.  
observed\_values = sample(c("Y","YA","A","O","OO"),size = n, replace = TRUE)  
  
number\_of\_obsers = table(observed\_values)  
data = data.frame(number\_of\_obsers)  
#####--The Expected outcome--#####  
expected\_values=c(0.2,0.2,0.2,0.2,0.2)   
  
#####-- Chi-square test--#####  
  
Chi\_test = chisq.test(data$Freq,p = expected\_values)  
print(Chi\_test)

##   
## Chi-squared test for given probabilities  
##   
## data: data$Freq  
## X-squared = 5.8, df = 4, p-value = 0.2146

Thus the p - value is greater than 0.05 which means we fail to reject the null hypothesis(Z. Bobbitt 2023 May 11), meaning the is an equal number of participation in the community by the age groups given. Majority of people tend to think that the assumption of the data being normally distributed must be followed to be able to use Goodness of fit chi - square test. But it does allow it to be used in normal data (Chen 2016).

# The Chi – square test of Independence.

This test determines if two categorical variables in a single sample are independent from each other or associated. Want to see if two variables are from the same sample. With the null hypothesis as “The variable of interests are independent”, and the alternative Hypothesis stating “The variables of interest are associated” (Franke, Ho, and Christie 2012).

An example is given where data is collected from one town and one wants to know if there is association between smoking and ending up being a criminal. this can be done by a survey by asking people if the smoke and if they commit a criminal activity regularly.

# The Chi – square test for homogeneity.

This test is used to test whether two or more independent samples differ in their distributions on a single variable of interest, it is used to compare two or more groups on a categorical outcome. With the null hypothesis as “The proportions between groups are the same “, and alternative hypothesis as “The proportions between groups are different”(Franke, Ho, and Christie 2012). in homogeneity post hoc tests need to be carried out if the null is rejected (Franke, Ho, and Christie 2012).

An example can be collecting data from two towns and one wants to know if the proportion of smokers in one town is equal to the other.

# Common misconceptions

The Chi- sqaure test of independence and for homogeneity are treated the same by many statisticians which leads to misinterpretations of the results and also leads to post hocc tests not being carried out after the is a test of significance (Franke, Ho, and Christie 2012).This misinterpretation is due to how the data is collected and sampled that is they have different sampling assumptions which is ignored. The difference is that for he test of independence data is collected on a single sample, and then compares two variables within that sample to determine the relationship between them, while the test for Homogeneity collects data on two or more distinct groups intentionally then compares the two samples on **a single variable of interest** to test whether the proportions differ between them (Franke, Ho, and Christie 2012).

# Conclusion

The assumptions must always be met for one to apply the chi - square test and there are techniques in place to prove if a data set meets the assumptions. To see if the is independence between observations one can check if there was random sampling (Zach Bobbitt 2021 Mrch 16) a new test must be used.The assumption of more than 5 observations will lead to using an alternative to the chi - square test like the fisher’s test, because if the chi - square test is used then the results will be misleading (Bower 2003). Chi square tests place more importance on the data collected.

# Reference

Bobbitt, Z. 2023 May 11. “Chi-Square Goodness of Fit Test: Definition, Formula, and Example.” 2023 May 11. <https://www.statology.org/chi-square-goodness-of-fit-test/>.

Bobbitt, Zach. 2021 Mrch 16. “What Is the Assumption of Independence in Statistics.” 2021 Mrch 16. <https://www.statology.org/assumption-of-independence/#:~:text=A%C2%A0two%20sample%20%EE%80%80t-testis%EE%80%81>.

Bower, Keith M. 2003. “When to Use Fisher’s Exact Test.” In *American Society for Quality, Six Sigma Forum Magazine*, 2:35–37. 4. American Society for Quality Milwaukee, WI, USA.

Chen, William W. S. 2016. “TWO EXAMPLES OF CHI-SQUARE GOODNESS-OF-FIT TEST.” *International Journal of Academic Research and Reflection* 2. <https://doi.org/ww.idpublications.org>.

Franke, Todd Michael, Timothy Ho, and Christina A. Christie. 2012. “The Chi-Square Test: Often Used and More Often Misinterpreted.” *American Journal of Evaluation* 33 (3): 448–58. <https://doi.org/10.1177/1098214011426594>.

ML, McHugh. 2013. “The Chi-Square Test of Independence.biochem Med (Zagreb),” 23:143–149. <http://dx.doi.org/10.11613/BM.2013.018>.