WWW.KRISHANPANDEY.COM

CHI SQUARE TEST

Dr. Krishan K. Pandey

2009

O P JINDAL GLOBAL UNIVERSITY

CHI-SQUARE (χ^2) ANALYSIS- INTRODUCTION

Consider the following decision situations:

Are all package designs equally preferred? 2) Are all brands equally preferred? 3) Is their any association between income level and brand preference? 4) Is their any association between family size and size of washing machine bought? 5) Are the attributes educational background and type of job chosen independent? The answers to these questions require the help of Chi-Square (χ^2) analysis. The first two questions can be unfolded using Chi-Square test of goodness of fit for a single variable while solution to questions 3, 4, and 5 need the help of Chi-Square test of independence in a contingency table. Please note that the variables involved in Chi-Square analysis are nominally scaled. Nominal data are also known by two names-categorical data and attribute data.

The symbol χ^2 used here is to denote the chi-square distribution whose value depends upon the number of degrees of freedom (d.f.). As we know, chi-square distribution is a skewed distribution particularly with smaller d.f. As the sample size and therefore the d.f. increases and becomes large, the χ^2 distribution approaches normality.

χ^2 tests are **nonparametric or distribution-free** in nature. This means that no assumption needs to be made about the form of the original population distribution from which the samples are drawn. Please note that all parametric tests make the assumption that the samples are drawn from a specified or assumed distribution such as the normal distribution.

CHI-SQUARE TEST-GOODNESS OF FIT

A number of marketing problems involve decision situations in which it is important for a marketing manager to know whether the pattern of frequencies that are observed fit well with the expected ones. The appropriate test is the χ^2 test of goodness of fit. The illustration given below will clarify the role of χ^2 in which only one categorical variable is involved.

Problem: In consumer marketing, a common problem that any marketing manager faces is the selection of appropriate colors for package design. Assume that a marketing manager wishes to compare five different colors of package design. He is interested in knowing which of the five the most preferred one is so that it can be introduced in the market. A random sample of 400 consumers reveals the following:

Package Color	preference by Consumers
Red	70
Blue	106
Green	80
Pink	70
Orange	74
Total	400

Do the consumer preferences for package colors show any significant difference?

Solution: If you look at the data, you may be tempted to infer that Blue is the most preferred color. Statistically, you have to find out whether this preference could have arisen due to chance. The appropriate test statistic is the χ^2 test of goodness of fit.

Null Hypothesis: All colors are equally preferred.

Alternative Hypothesis: They are not equally preferred

Package Color	**Observed Frequencies (O)**	**Expected Frequencies (E)**	$(O-E)^2$	$\chi^2 = \sum\left(\frac{(O-E)^2}{E}\right)$
Red	**70**	**80**	**100**	**1.250**
Blue	**106**	**80**	**676**	**8.450**
Green	**80**	**80**	**0**	**0.000**
Pink	**70**	**80**	**100**	**1.250**
Orange	**74**	**80**	**36**	**0.450**
Total	**400**	**400**		**11.400**

Please note that under the null hypothesis of equal preference for all colors being true, the expected frequencies for all the colors will be equal to 80. Applying the formula

$$\chi^2 = \sum\left(\frac{(O-E)^2}{E}\right),$$

we get the computed value of chi-square (χ^2) = 11.400

The critical value of χ^2 at 5% level of significance for 4 degrees of freedom is 9.488. So, the null hypothesis is rejected. The inference is that all colors are not equally preferred by the consumers. In particular, Blue is the most preferred one. The marketing manager can introduce blue color package in the market.

CHI-SQUARE TEST OF INDEPENDENCE

The goodness-of-fit test discussed above is appropriate for situations that involve one categorical variable. If there are two categorical variables, and our interest is to examine whether these two variables are associated with each other, the chi-square (χ^2) test of independence is the correct tool to use. This test is very popular in analyzing cross-tabulations in which an investigator is keen to find out whether the two attributes of interest have any relationship with each other.

The cross-tabulation is popularly called by the term "contingency table". It contains frequency data that correspond to the categorical variables in the row and column. The marginal totals of the rows and columns are used to calculate the expected frequencies that will be part of the computation of the χ^2 statistic.

Problem: A marketing firm producing detergents is interested in studying the consumer behavior in the context of purchase decision of detergents in a specific market. This company is a major player in the detergent market that is characterized by intense competition. It would like to know in particular whether the income level of the consumers influence their choice of the brand. Currently there are four brands in the market. Brand 1 and Brand 2 are the premium brands while Brand 3 and Brand 4 are the economy brands.

A representative stratified random sampling procedure was adopted covering the entire market using income as the basis of selection. The categories that were used in classifying income level are: Lower, Middle, Upper Middle and High. A sample of 600 consumers participated in this study. The following data emerged from the study. {Cross Tabulation of Income versus Brand chosen (Figures in the cells represent number of consumers)}

	Brands				
	Brand1	**Brand2**	**Brand3**	**Brand4**	**Total**
Income					
Lower	**25**	**15**	**55**	**65**	**160**
Middle	**30**	**25**	**35**	**30**	**120**
Upper Middle	**50**	**55**	**20**	**22**	**147**
Upper	**60**	**80**	**15**	**18**	**173**
Total	**165**	**175**	**125**	**135**	**600**

Analyze the cross-tabulation data above using chi-square test of independence and draw your conclusions.

Solution:

Null Hypothesis: There is no association between the brand preference and income level (These two attributes are independent).

A**lternative Hypothesis**: There is association between brand preference and income level (These two attributes are dependent).

Let us take a level of significance of 5%.

In order to calculate the χ^2 value, you need to work out the expected frequency in each cell in the contingency table. In our example, there are 4 rows and 4 columns amounting to 16 elements. There will be 16 expected frequencies.

Observed Frequencies (These are actual frequencies observed in the survey)

	Brands				
	Brand1	**Brand2**	**Brand3**	**Brand4**	**Total**
Income					
Lower	**25**	**15**	**55**	**65**	**160**
Middle	**30**	**25**	**35**	**30**	**120**
Upper Middle	**50**	**55**	**20**	**22**	**147**
Upper	**60**	**80**	**15**	**18**	**173**
Total	**165**	**175**	**125**	**135**	**600**

Expected Frequencies (These are calculated on the assumption of the null hypothesis being true: That is, income level and brand preference are independent)

	Brands				
	Brand1	**Brand2**	**Brand3**	**Brand4**	**Total**
Income					
Lower	**44.000**	**46.667**	**33.333**	**36.000**	**160.000**
Middle	**33.000**	**35.000**	**25.000**	**27.000**	**120.000**
Upper Middle	**40.425**	**42.875**	**30.625**	**33.075**	**147.000**
Upper	**47.575**	**50.458**	**36.042**	**38.925**	**173.000**
Total	**165.000**	**175.000**	**125.000**	**135.000**	**600.000**

Note: The fractional expected frequencies are retained for the purpose of accuracy. Do not round them.

Calculation: Compute

$$\chi^2 = \sum \left(\frac{(O - E)^2}{E} \right)$$

There are 16 observed frequencies (O) and 16 expected frequencies (E). As in the case of the goodness of fit, calculate this χ^2 value. In our case, the computed χ^2 =131.76 as shown below: Each cell in the table below shows $(O-E)^2/(E)$

	Brand1	Brand2	Brand3	Brand4
Income				
Lower	**8.20**	**21.49**	**14.08**	**23.36**
Middle	**0.27**	**2.86**	**4.00**	**0.33**
Upper Middle	**2.27**	**3.43**	**3.69**	**3.71**
Upper	**3.24**	**17.30**	**12.28**	**11.25**

And there are 16 such cells. Adding all these 16 values, we get χ^2 =131.76

The critical value of χ^2 depends on the degrees of freedom. The degrees of freedom = (the number of rows-1) multiplied by (the number of colums-1) in any contingency table. In our case, there are 4 rows and 4 columns. So the degrees of freedom =(4-1). (4-1) =9. At 5% level of significance, critical χ^2 for 9 d.f = 16.92. Therefore reject the null hypothesis and accept the alternative hypothesis.

The inference is that brand preference is highly associated with income level. Thus, the choice of the brand depends on the income strata. Consumers in different income strata prefer different brands. Specifically, consumers in upper middle and upper income group prefer premium brands while consumers in lower income and middle-income category prefer economy brands. The company should develop suitable strategies to position its detergent products. In the marketplace, it should position economy brands to lower and middle-income category and premium brands to upper middle and upper income category.