NON PARAMETRIC

CHI- SQUARE TEST

Karl Pearson developed a statistical test called the χ^2 -test, to test the significance of the difference between observed and expected frequencies. This if very widely and commonly used in biological sciences, especially, genetics. Certain terms, such as Null-hypothesis, level of significance, degrees of freedom, etc are simply defined here.

The Chi-square test, written as χ^2 - test, is a useful measure of comparing experimentally obtained results with those expected theoretically and based on the hypothesis. It is used as a test statistic in testing a hypothesis that provides a set of theoretical frequencies with which observed frequencies are compared. In general Chi-square test is applied to those problems in which we study whether the frequency with which a given even has occurred, is significantly different from the one as expected theoretically. The measure of Chi-square enables us to find out the degree of discrepancy between observed frequencies and theoretical frequencies and thus to determine whether the discrepancy so obtained between observed frequencies and theoretical frequencies is due to error of sampling or due to chance.

The Chi-square is computed on the basis of frequencies in a sample and thus the value of Chi-square so obtained is a statistic. Chi-square is not a parameter as its value is not derived from the observations in a population. Hence Chi-square test is a Non- Parametric test. Chi-square test is not concerned with any population distribution and its observations.

The χ^2 - test was first used in testing statistical hypothesis by Karl Pearson in the year 1900. It is defined as

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

Where = O_i = Observed frequency of i^{th} event

E_i = Expected frequency of i^{th} event

We require the following steps to calculate χ^2

Step 1. Calculate all expected frequencies ,i.e. E_i for all values of i=1,2,….,n

Step2. Take the difference between each observed frequency O_i and the

corresponding expected frequency E_i for each value of i , i.e.,find (O_i - E_i)

Step3. Square the difference for each value of i,i.e., calculate (O_i - E_i) ^{2} for all

values of i = 1,2,3,……n.

Step4. Divide each square difference by the corresponding expected frequency , i.e., calculate $\frac{(O_i - E_i)^2}{E_i}$ for all values of i = 1,2,3,……n

Step 5. Add all these quotients obtained in STEP 4 , then

$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$ is the required value of Chi-square.

It should be noted that

(a) The value of χ^2 is always positive as each pair is squared one.

(b) χ^2 will be zero if each pair is zero and it may assume any value extending to infinity, when the difference between the observed frequency and expected frequency in each pair is unequal .Thus χ^2 lies between 0 and ∞.

(c) The significance test on χ^2 is always based on One Tailed Test of the right hand side of standard normal curve as χ^2 is always non-negative.

(d) As χ^2 is a statistics and not a parameter, so it does not involve any assumption about the form of original distribution from which the observations have come.

DEGREES OF FREEDOM

The number of data that are given in the form of a series of variables in a row or column or the number of frequencies that are put in cells in a contingency table, which can be calculated independently is called the degrees of freedom and is denoted by v.

Case I If the data is given in the form of a series of variables in a row or column, then the degrees of freedom = (number of items in the series) – 1, i.e., v =n -1, where n is the number of variables in the series in a row or column.

Case II When the number of frequencies are put in cells in a contingency table, the

degrees of freedom will be the product of (number of rows less one)and the

(number of columns less one) i.e., v = (R-1) (C-1). Where R is the number of rows and C is the number of columns.

PROPERTIES OF χ^2 -DISTRIBUTION

1. Chi-square curve is always positively skewed.
2. The mean of χ^2 distribution is the number of degrees of freedom
3. The standard deviation of χ^2 -distribution = $\sqrt{2v}$ where v is the degrees of Freedom.
4. Chi-Square values increases with the increase in degrees of freedom.
5. The value of χ^2 lies between zero and infinity ,i.e., $0 \leq \chi^2 < \infty$
6. The sum of two χ^2 distribution is again a χ^2 distribution, i.e., are two independent and they have a χ^2 distribution with n_1 and n_2 degrees of freedom respectively ,then($\chi_1^2 + \chi_2^2$) is also χ^2 distribution with (n_1+n_2) degrees of freedom.
7. For different degrees of freedom, the shape of the curve will be different
9. Its shape depends on the degree of freedom but it is not a symmetrical distribution.

Chi square test are becoming more popular in applied research because of its multifarious application and uses. Before discussing its application, let us define what the chi square is?

The square of a standard normal variate is known as a chi square variate with 1 degree of freedom. Thus if $x \sim N(\mu, \sigma^2)$

then $Z = \prec \frac{x-\mu}{\sigma} \sim N(0,1)$

and $Z^2 = \left(\frac{x-\mu}{\sigma}\right)^2$ is a chi square variate with 1 d.f.

In general if X_i (i = 1,2 ,….,n) are n independent normal variates with mean μ_i and variance σ_i (i = 1,2,…,n) then

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2$$

Is a chi-sqaure variate with n d.f. Probability curve of the chi square changes with degree of freedom. Further χ^2 axis is an asymptotic to the curve.

Conditions for the Validity of χ^2 Test

The χ^2 test would be valid if the following conditions are satisfied :

I . The sample observations should be independent ;

II Sum of the observed frequencies should be equal to the sum of the expected Frequencies.

III The total frequencies (N) should be large enough, say greater than 50.

IV. No theoretical cell frequencies should be less than 5 because in that case χ^2 can not maintain the continuity. In case of any theoretical cell frequency less then 5, it should be pooled with the preceding or succeeding frequency so that the pooled frequency is more than 5 and accordingly loss in the d.f. due to pooling should be adjusted.

Critical Values

It may be observed from the chi square Table that theoretical values of χ^2 increases as n (d.f) increases and level of significance decreases. Let $\chi^2_\alpha(n)$ denotes the value of tabulated chi square for n d.f. such that the area to the right of this point is α , i.e.

$$P\left[\chi^2 > \chi^2_\alpha(n)\right] = \alpha$$

Application of Chi-Square (χ^2)

Chi-square distribution has a large number of applications in applied research but a few which are widely used are enumerated below:

I : To test the equal occurrence hypothesis.

II : To test the independence of attributes.

III . To compare the attitude in two different groups.

Equal Occurrence Test

INDEPENDENCE OF ATTRIBUTES

Example 12.3 Following are the frequencies of preferences of male and female students towards sports.

Sex	prefer sports	Do not prefer sports
Male	55	20
Female	20	30

Test whether sex is related with the preference of sports.

Solution

OBJECTIVE

To test whether preferences of sports is sex based ? or in other words whether the responses of male towards sports are same in comparison to that of female ?

HYPOTHESIS

H_O : There is no association between sex and sport preference

H_1 : There is an association between sex and sport preference

LEVEL OF SIGNIFICANCE

$$\alpha = .05$$

CALCULATION OF TEST STATISTIC

Observed frequencies (fo)

Sex	Prefer Sports	Do not prefer sports	Total
Male	55	20	75
female	20	30	50
Total	75	50	N = 125

Thus expected frequencies are shown in the following table

Expected frequencies (Fe)

Sex	Prefer Sports	Do not prefer sports
Male	$\frac{75 \times 75}{125} = 45$	$\frac{50 \times 75}{125} = 30$
female	$\frac{75 \times 50}{125} = 30$	$\frac{50 \times 50}{125} = 20$

d.f. = (r-1) (c -1) = (2-1) (2-1) =1

Tabulated $\chi^2_{0.05}(1) = 3.84$

STATISTICAL DECISION

Since calculated χ^2 is greater than tabulated χ^2 thus null hypothesis of no association between sex and sports preference may be rejected an .05 level of significance.

INFERENCE

There is an association between sex and preference towards sports. In other words it could be inferred that the males response towards preference of sports is different than that of females.

Problem

A public opinion poll surveyed a simple random sample of 1000 voters. Respondents were classified by gender (male or female) and by voting preference (Republican, Democrat, or Independent). Results are shown in the contingency table below.

	Voting Preferences			Row total
	Rep	Dem	Ind	
Male	200	150	50	400
Female	250	300	50	600
Column total	450	450	100	1000

Is there a gender gap? Do the men's voting preferences differ significantly from the women's preferences? Use a 0.05 level of significance.

Solution

The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results. We work through those steps below:

Hypothesis

H_o: Gender and voting preferences are independent.

H_a: Gender and voting preferences are not independent.

Level of significance

For this analysis, the significance level is 0.05.

Analyze sample data. Applying the chi-square test for independence to sample data, we compute the degrees of freedom, the expected frequency counts, and the chi-square test statistic. Based on the chi-square statistic and the degrees of freedom, we determine the P-value.

$$DF = (r - 1) * (c - 1) = (2 - 1) * (3 - 1) = 2$$

$$E_{r,c} = (n_r * n_c) / n$$

$$E_{1,1} = (400 * 450) / 1000 = 180000/1000 = 180$$

$$E_{1,2} = (400 * 450) / 1000 = 180000/1000 = 180$$

$$E_{1,3} = (400 * 100) / 1000 = 40000/1000 = 40$$

$$E_{2,1} = (600 * 450) / 1000 = 270000/1000 = 270$$

$$E_{2,2} = (600 * 450) / 1000 = 270000/1000 = 270$$

$$E_{2,3} = (600 * 100) / 1000 = 60000/1000 = 60$$

- $X^2 = \Sigma [(O_{r,c} - E_{r,c})^2 / E_{r,c}]$

$$\chi^2 = (200 - 180)^2/180 + (150 - 180)^2/180 + (50 - 40)^2/40$$

$$+ (250 - 270)^2/270 + (300 - 270)^2/270 + (50 - 60)^2/60$$

$$\chi^2 = 400/180 + 900/180 + 100/40 + 400/270 + 900/270 + 100/60$$

$$\chi^2 = 2.22 + 5.00 + 2.50 + 1.48 + 3.33 + 1.67 = 16.2$$

where DF is the degrees of freedom, r is the number of levels of gender, c is the number of levels of the voting preference, n_r is the number of observations from level *r* of gender, n_c is the number of observations from level *c*of voting preference, n is the number of observations in the sample, $E_{r,c}$ is the expected frequency count when gender is level *r* and voting preference is level *c*, and $O_{r,c}$ is the observed frequency count when gender is level *r* voting preference is level *c*.

The P-value is the probability that a chi-square statistic having 2 degrees of freedom is more extreme than 16.2.

To find $P(\chi^2 > 16.2) = 0.0003$.

- **Interpret results**. Since the P-value (0.0003) is less than the significance level (0.05), we cannot accept the null hypothesis. Thus, we conclude that there is a relationship between gender and voting preference.