

UNIT 110 – IPSC - Chi Square Test

The statistical procedures are suitable only for numerical variables. The chi square distribution is a theoretical or mathematical distribution which is extensively applicable in statistical work. The chi-square test is an important test among various tests of significance developed by statisticians. It was developed by Karl Pearson in 1900. Chi square test is a nonparametric test not based on any assumption or distribution of any variable. The term 'chi square' (pronounced with a hard 'ch') is used because the Greek letter χ is used to define this distribution. It will be seen that the elements on which this distribution is based are squared, therefore the symbol χ^2 is used to signify the distribution.



The chi-square (χ^2) test is an important test which can be used to assess a relationship between two categorical variables. It is one example of a nonparametric test. A chi-squared test is any statistical hypothesis test in which the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true.

Chi Square Formula

$$\chi^2 = \sum \frac{(\text{Observed Value} - \text{Expected Value})^2}{(\text{Expected Value})}$$

In general, the chi-square test is used to determine whether there

is a significant difference between the expected frequencies and the observed frequencies in one or more categories.

Requirements for Chi Square Test

1. Quantitative data.
2. One or more categories.
3. Independent observations.
4. Adequate sample size (at least 10).
5. Simple random sample.
6. Data in frequency form.
7. All observations must be used.

Characteristics of a chi square test: Following are the main characteristics of chi square test:

1. This test is based on frequencies and not on the parameters like mean and standard deviation.
2. The test is used for testing the hypothesis and is not useful for estimation.
3. This test can also be applied to a complex contingency table with several classes and as such is a very useful test in research work.
4. This test is an important non-parametric test as no rigid assumptions are necessary in regard to the type of population, no need of parameter values and relatively less mathematical details are involved.

The Chi Square Statistic: The χ^2 statistic seems different from the other statistics which have been used in the previous hypotheses tests. It also has some similarity to the theoretical chi square distribution. For both the goodness of fit test and the test of independence, the chi square statistic is the same. For both of these tests, all the categories into which the data have been divided are used. The data obtained from the sample are termed as the observed numbers of cases. These are the frequencies of occurrence for each class into which the data have been grouped. In the chi square tests, the null hypothesis makes a statement concerning how many cases are to be expected in each category if this hypothesis is correct. Another application of chi square test is test of homogeneity.

The chi square goodness of fit test: This test begins by hypothesizing that the distribution of a variable behaves in a particular manner. The goodness of fit, abbreviated as GOF tests measure the compatibility of random sample with a theoretical probability distribution function. Common method comprises of defining a test statistic which is some function of the data measuring the distance between the hypothesis and the data, and then calculating the probability of obtaining data which have a still larger value of this test statistic than the value observed, assuming the hypothesis is true. This probability is called the confidence level. The chi-square goodness of fit test is suitable in the following conditions:

1. The sampling method is simple random sampling.
2. The population is at least 10 times as large as the sample.
3. The variable under study is categorical.
4. The expected value of the number of sample observations in each level of the variable is at least 5.

It is established that researcher must use the chi-square test of goodness-of-fit when they have one nominal variable with two or more values. They compare the observed counts of observations

in each category with the expected counts, which you calculate using some kind of theoretical expectation. If the expected number of observations in any category is too small, the chi-square test may give inaccurate results, and researcher should use an exact test instead.

This approach include four steps:

1. State the hypotheses,
2. Formulate an analysis plan,
3. Analyse sample data
4. Interpret results.

The test of independence: Test facilitates researchers to explain whether or not two attributes are associated. The Chi-Square Test of Independence is also called Pearson's Chi-Square. The chi-square test of independence is a nonparametric statistical analysis method often used in experimental work where the data consist in frequencies or 'counts'. The general use of the test is to assess the probability of association or independence of facts.

Chi-Square Test for Independence is used in following conditions:

1. The sampling method is simple random sampling.
2. Each population is at least 10 times as large as its respective sample.
3. The variables under study are each categorical.
4. If sample data are displayed in a contingency table, the expected frequency count for each cell of the table is at least

Test of homogeneity: This test can also be used to test whether the occurrence of events follow uniformity or not. Chi-square test of homogeneity is used to a single categorical variable from two different populations. It is used to determine whether frequency counts are distributed identically across different populations. Conditions for using Chi-Square Test for Homogeneity:

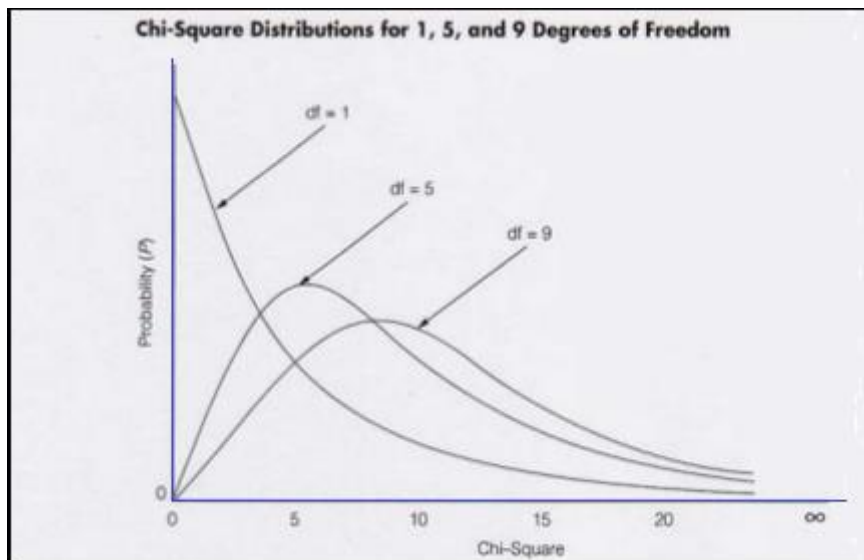
1. For each population, the sampling method is simple random sampling.
2. Each population is at least 10 times as large as its respective sample.
3. The variable under study is categorical.
4. If sample data are displayed in a contingency table (Populations x Category levels), the expected frequency count for each cell of the table is at least

The test for Homogeneity is evaluating the equality of several populations of categorical data.

The homogeneity chi-square test statistics is computed exactly the same as the test for independence using contingency table when determining the independence of characteristics chi-square statistics.

Main difference between the test for independence and the homogeneity test is the stating of the null hypothesis. Homogeneity tests a null hypothesis asserting that various populations are homogeneous or equal with respect to some characteristics of interest against an alternate hypothesis claiming that they are not.

Calculation of Chi Square Each entry in the summation is termed as "The observed minus the expected, squared, and divided by the expected." The chi square value for the test as a whole is "The sum of the observed minus the expected, squared, and divided by the expected."



Limitations of the Chi-Square Test: Though chi square test is very useful in statistics but it also has some drawbacks:

1. The chi-square test does not give us much information about the strength of the relationship or its substantive significance in the population.
2. The chi-square test is sensitive to sample size. The size of the calculated chi-square is directly proportional to the size of the sample, independent of the strength of the relationship between the variables.
3. The chi-square test is also sensitive to small expected frequencies in one or more of the cells in the table.

To summarize, Chi-square is a statistical test usually used to compare or observed data by data analyst to obtain requisite results according to a specific hypothesis. Among two different types of chi square tests, the goodness of fit test allows the researcher to hypothesize that a distribution behaves in a particular manner. The chi square

test is used to decide whether the observed distribution conforms to the model hypothesized by the scholar in the null hypothesis. In the test for independence, the chi square statistic and distribution are used to test whether there is a relationship between the two variables or not. For each of the two tests, the chi square statistic and distribution are the same. Though, the two tests are relatively different types of hypothesis tests, one investigating only a single variable, and the other examining the relationship between two variables.

Anova Test

The Analysis Of Variance, commonly known as the ANOVA, can be used in conditions where there are more than two groups. ANOVA was developed in the 1920's by R.A. Fisher. Analysis of Variance is a statistical technique to determine the extent to which the independent variable(s) impact the dependent variables in a regression analysis. This assists to identify the factors that influence the given set of data. The variance in a data set can be attributed mainly to-

Random factors that do not hold any statistical significance in the analysis.

Systematic factors that are important to understand its statistical influence.

ANOVA is based on the comparison of the average value of the variance among groups relative to variance within groups (Random Error). When ANOVA test is executed, it is possible to identify the systematic factors that are statistically contributing to the data set's variability. In common way, ANOVA provides a statistical test of whether or not the means of several groups are all equal, and therefore generalizes t-test to more than two groups. ANOVA tests are beneficial to test three or more means (groups or variables) for statistical significance.

The test for ANOVA is the ANOVA F-test. The main objective of ANOVA is to test for significant differences between means. Elementary Concepts provides a brief introduction to the basics of statistical significance testing. ANOVA is used to test for differences among several means without increasing the Type I error rate. This test uses data from all groups to estimate standard errors, which can increase the power of the analysis.

The Null hypothesis for ANOVA is that the means for all groups are equal:
 $H_0 : \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$

The Alternative hypothesis for ANOVA is that at least two of the means are not equal.

1. Rejecting the null hypothesis doesn't require that ALL means are significantly different from each other.
2. If at least two of the means are significantly different the null hypothesis is rejected.

There are some assumptions which must be considered before using ANOVA:

1. The observations are from a random sample and they are independent from each other
 2. The observations are assumed to be normally distributed within each group.
-
1. ANOVA is still suitable if this assumption is not met but the sample size in each group is large (> 30)
 2. The variances are approximately equal between groups. If the ratio of the largest SD / smallest SD < 2 , this assumption is considered to be met.

Steps in ANOVA:

1. Description of data
2. Assumption: Along with the assumptions, we represent the model for each design we discuss.
3. Hypothesis
4. Test statistic
5. Distribution of test statistic
6. Decision rule
7. Calculation of test statistic: The results of the arithmetic calculations is summarized in a table called the analysis of variance (ANOVA) table. The entries in the table make it easy to evaluate the results of the analysis.
8. Statistical decision
9. Conclusion
10. Determination of p value

Types of ANOVA

One-Way ANOVA: It is the simplest type of ANOVA, in which only one source of variation, or factor, is investigated. One way ANOVA is common statistical technique to determine if differences exist between two or more "groups". One-way ANOVA is an extension to three or more samples of the t-test procedure for use with two independent samples. This is done when there is only one qualitative variable which denotes the groups and only one measurement variable (quantitative). For a one-way ANOVA, the observations are divided into mutually exclusive categories, giving the one-way classification. The purpose of one way Anova is to validate whether the data collected from different sources converge on a common mean. Assumptions of the One-Way ANOVA are;

1. The data are randomly sampled
2. The variances of the populations are equal
3. The distribution of scores in each population are normal in shape

The one way Anova uses only one category of defining characteristics to carry out its procedure.

Two-Way ANOVA: Two way ANOVA is applied where an experiment has a quantitative outcome and two categorical explanatory variables that are defined in such a way that each experimental unit can be exposed to any combination of one level of one explanatory variable and one level of the other explanatory variable. Two-Way ANOVA has 2 independent variables (factors) and each can have multiple conditions. The aim of the two way Anova is to verify whether the data collected from different sources coverage on a common mean based on two categories of defining characteristics.

In two-way ANOVA, the error model is the normal one of Normal distribution with equal variance for all subjects that share levels of both of the explanatory variables. It is established in quantitative studies that Two-way ANOVA is suitable analysis method for a study with a quantitative outcome and two (or more) categorical explanatory variables. The usual assumptions of Normality, equal variance, and independent errors apply.

Advantages of a two-way ANOVA model are as follows:

1. The 2-way ANOVA breaks down the fit part of the model between each of the main components (the 2 factors) and an interaction effect. The interaction cannot be tested with a series of one-way ANOVAs.
2. It usually have a smaller total sample size, since researchers study two things at once.
3. The 2-way ANOVA removes some of the random variability.
4. Researcher can look at interactions between factors (a significant interaction means 1. the effect of one variable changes depending on the level of the other factor).

To summarize, Analysis of variance is a method for testing the hypothesis that there is no difference between two or more population means. ANOVA is used in the analysis of comparative experiments, those in which only the difference in outcomes is of interest.