**Notes – Ch12 Tests of Goodness of Fit and Independence**

The chi-square χ2 (rhymes with “sky”) distribution is a sampling distribution. The chi-square distribution is a *family* of probability distributions. As with the *t* distribution, the specific member depends on a value for the number of degrees of freedom. The chi-square distribution is skewed positively, but as *df* increases, it approaches the shape of the normal distribution

**Chi-square test can be used for:**

1. **Test for independence:** When data involve two nominal-scale variables, with each variable having two or more categories, chi-square analysis can be used to test whether or not there might be some relationship between the variables. For example, a campaign manager might be interested in whether party affiliation could be related to educational level, or an insurance underwriter may wish to determine whether different occupational groups tend to suffer different kinds of on-the-job injuries. The purpose of chi-square analysis in such applications is not to identify the exact nature of a relationship between nominal variables; the goal of this technique is simply to test whether or not the variables could be independent of each other.
2. **Goodness of fit:** Goodness-of-fit test is a statistical test for determining whether there is a significant difference between an observed frequency distribution and a theoretical probability distribution hypothesized to describe the observed distribution. Chi-square analysis is applied for the purpose of examining whether sample data could have been drawn from a population having a specified probability distribution. For example, we may wish to determine from sample data

(1) whether a set of data could have come from a population having a multinomial distribution, A multinomial distribution can be thought of extension of the binomial distribution to the case of three or more categories of outcomes. On each trial of a multinomial experiment, one and only one of the outcomes occur. Each trial of the experiment is assumed to be independent, and the probabilities of the outcomes remain the same for each trial.

(2) whether a set of data could have come from a population having a Poisson distribution,

(3) whether a set of data could have come from a population having a normal distribution.

1. **Test for independence:** The starting point for the chi-square test of variable independence is the contingency table.

**Contingency Table:** A table having n rows and m columns. Each row corresponds to a level of one variable, each column to a level of another variable. Entries in the body of the table are the frequencies with which each variable combination occurred.

The null and alternative hypotheses are

*H*0:  The variables are independent of each other.

*H*a:  The variables are not independent of each other.

There will be a table of observed frequencies and a table of expected frequencies, and the amount of disparity between these tables will be compared during the calculation of the test statistic. The observed frequencies will reflect a cross-classification for members of a single sample, and the table of expected frequencies will be constructed under the assumption that the null hypothesis is true. The expected frequencies for contingency tables under assumption of independence is given by:

Test statistics for Independence is given by:

fij= observed frequency for the contingency table category in row i and column j

eij= observed frequency for the contingency table category in row i and column j

The p-value is compared to a given level of significance using (n-1)(m-1) degrees of freedom where n,m are no of rows,columns.

Rejection Rule :

|  |  |
| --- | --- |
| p-value approach: | if then reject H0 |
| Critical value approach: | If then reject H0 |

1. **Goodness of fit for multinomial distribution**

The null and alternative hypotheses are

*H*0: The population follows a multinomial distribution with specified probabilities for each of the *k* categories

*H*a: The population does not follow a multinomial distribution with the specified probabilities for each of the *k* categories

Select a random sample and record the observed frequencies *fi* for each category.

Assume the null hypothesis is true and determine the expected frequency *ei* in each category by multiplying the category probability by the sample size.

*α* is the level of significance for the test and there are k – 1 degrees of freedom where k is no of categories.

Rejection Rule :

|  |  |
| --- | --- |
| p-value approach: | if then reject H0 |
| Critical value approach: | If then reject H0 |

1. **Goodness of fit for Poisson distribution**

*When the expected number in some category is less than five, the assumptions for the chi-square are not satisfied. When this happens, adjacent categories can be combined to increase the expected number to five.*

The null and alternative hypotheses are

*H*0: The population has a Poisson distribution

*H*a: The population does not have a Poisson distribution

Select a random sample and

a. Record the observed frequency *fi* for each value of the Poisson random  variable.

b. Compute the mean number of occurrences *μ*.

Compute the expected frequency of occurrences *ei* for each value of the Poisson random variable. Multiply the sample size by the Poisson probability of occurrence for each value of the Poisson random variable. If there are fewer than five expected occurrences for some values, combine adjacent values and reduce the number of categories as necessary.

*α* is the level of significance for the test and there are k – p - 1 degrees of freedom where k is no of categories and p is no of parameters. Because the sample data were used to estimate the mean of the Poisson distribution, p = 1.(One parameter i.e. mean)

Rejection Rule :

|  |  |
| --- | --- |
| p-value approach: | if then reject H0 |
| Critical value approach: | If then reject H0 |

1. **Goodness of fit for Normal distribution**

When the expected number in some category is less than five, the assumptions for the chi-square are not satisfied. When this happens, adjacent categories can be combined to increase the expected number to five.

The null and alternative hypotheses are

*H*0: The population has a normal distribution

*H*a: The population does not have a normal distribution

Select a random sample and

a. Compute the sample mean and sample standard deviation.

b.Define intervals of values so that the expected frequency is atleast five for

each interval. Using equal probability intervals is a good approach.

**c.** Record the observed frequency of data values *fi* in each interval defined.

Compute the expected number of occurrences *ei* for each interval of values defined in step (b). Multiply the sample size by the probability of a normal random variable being in the interval.

*α* is the level of significance for the test and there are k – p - 1 degrees of freedom where k is no of categories and p is no of parameters. p = 2 as mean and standard deviation are the 2 parameters used for normal distribution.

Rejection Rule :

|  |  |
| --- | --- |
| p-value approach: | if then reject H0 |
| Critical value approach: | If then reject H0 |