Goodness of fit

The goodness of fit test involves a comparison of the observed frequency of occurrence of classes with that predicted by a theoretical model. Suppose there are n classes with *observed frequencies* O_1, O_2, ..., O_n, and corresponding *expected frequencies* E_1, E_2, ..., E_n. The expected frequency is the average number or expected value when the hypothesis is true and is simply calculated as n multiplied by the hypothesized population proportion. The statistics

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

has a distribution that is distributed *approximately* as χ^2 with n -1 degrees of freedom. This approximation becomes better as n increases. If parameters from the data are used to calculate the expected distributions the degrees of freedom of the χ^2 will be n –1–p; where p is the number of parameteres estimated.

That is goodness of fit test is used to test

- For testing significance of patterns in qualitative data
- Test statistic is based on counts that represent the number of items that fall in each category
- Test statistics measures the agreement between actual counts and expected counts. The chi-square distribution can be used to see whether or not an observed counts agree with an expected counts.

Hypothesis:

Null hypothesis: Assumes that the given data follows a specific distribution (binomial, Normal etc).

Alternative hypothesis: Assumes that the given data does not follows a specific distribution

Procedure:

First we have to calculate the expected value or the expected frequency E of the fit model after fitting the distribution. The given frequency is the observed frequency. After calculating the expected value, we will apply the following formula to calculate the value of the Chi-Square

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where O refers to the observed frequencies and E for the expected frequencies. This statistics has Chi-Square distribution with df = number of observation- number of parameter estimated-1. If the hypothesis directly specifies the theoretical frequencies or the distribution df = number of observation- 1.

Decision criteria is Reject H_0 if $\chi^2 \geq \chi^2_\alpha$.

The tabled corresponding to the calculated degrees of freedom for a given α.

Contingency tables

For more than one variable, data can be conveniently represented by two-way tables called *contingency* table. These tables are useful to test if two classification criteria are independent (test of independence) and if two samples belong to the same population in relation to one-classification criteria (test of homogeneity). These tests are based on the principle that if two events are independent, the probability of their occurring together can be computed as the product of their separate probabilities.

ASSOCIATION OF ATTRIBUTE

- The Degree of relationship between two or more attributes is called association

Positive and negative association :

- Two attributes are said to positively associated if the presence of one result in the presence of the other. Eg. Education and employment
- Two attributes are said to be negatively associated if the presence of one result in the absence of the other. Eg. Immunization and Infection.

Chi-Square Test of Independence

The Chi-Square test of Independence is used to determine if there is a significant relationship between two nominal (categorical) variables. The frequency of one nominal variable is compared with different values of the second nominal variable. The data can be displayed in an n*m contingency table, where n is the row and m is the column. For example, a researcher wants to examine the relationship between gender (male vs. female) and empathy (high vs. low). The chi-square test of independence can be used to examine this relationship. If the null hypothesis is accepted there would be no relationship between gender and empathy. If the null hypotheses is rejected the implication would be that there is a relationship between gender and empathy (e.g. females tend to score higher on empathy and males tend to score lower on empathy).

Hypothesis:

- **Null hypothesis:** Assumes that there is no association between the two variables.
- **Alternative hypothesis:** Assumes that there is an association between the two variables.

Procedure:

First we have to calculate the expected value of the two nominal variables. We can calculate the expected value of the two nominal variables by using this formula:

$$E_i = \frac{R_i \times C_i}{N}$$

Where R_i is the raw total and C_i is the column total of i^{th} cell.

After calculating the expected value, we will apply the following formula to calculate the value of the Chi-Square test of Independence:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Degree of freedom is calculated by using the following formula:
df = (r-1)(c-1)
Where r = number of rows and c = number of columns.

The interpretation

In every χ2-test the calculated χ^2 value will either be (i) less than or equal to the critical χ^2 value OR (ii) greater that the critical χ^2 value.

• If calculated $\chi^2 \leq$ critical χ^2, then we conclude that there is *no statistically significant difference* between the two distributions. That is, the observed results are not significantly different from the expected results, and the numerical difference between observed and expected can be attributed to chance.

• If calculated $\chi^2 >$ critical χ^2, then we conclude that there **is** a *statistically significant difference* between the two distributions. That is, the observed results **are** significantly different from the expected results, and the numerical difference between observed and expected can **not** be attributed to chance. That means that the difference found is due to some other factor. This test won't identify that other factor, only that there is some factor other than chance responsible for the difference between the two distributions.