Practical Session 5 : chisquare tests

Goodness-of-fit tests

Teachers conducted a market resarch survey by asking computer science students and staff what type of chocolate they preferred from a list of milk chocolate, dark chocolate, white chocolate etc. They want to use a chi-squared goodness-of-fit test to check whether the preferences of computer scientists match those of french population as a whole.

Here are the data

	Dark	Milk	White	Other
Whole population	50000	60000	100000	35000
Computer scientists	25	30	60	15

1. Define two data frames corresponding to these two populations. Print the two dataframes
2. Chi-squared tests are based on the so-called chi-squared statistic. You calculate the chi-squared statistic with the following formula :

$$\sum \frac{(observed - expected)^2}{expected}$$

   Calculate this test statistic using the formula with Python
3. To determine whether the result is significant, in the chi-square test we have to compare the chi-square test statistic to a critical value based on the chi-square distribution. Use the scipy library to find the critical value for 95% confidence level and check the p-value. Do you accept or reject the null hypothesis of adequation of the two distribution ? Carry out a chi-squared goodness-of-fit test automatically using the scipy function `scipy.stats.chisquare()`

Independence test

We want uncover the factors that lead to employee attrition and explore important questions such as 'show me a breakdown of distance from home by job role and attrition' or 'compare average monthly income by education and attrition'. This is a fictional data set created by IBM data scientists.

1. Import the dataset in Python. Check out the number of employees and the number of attributes. Is there any missing values ?
2. We want to examine if there is a relationship between 'Attrition' and 'JobSatisfaction'.
   (i) Count for the two categories of 'Attrition'.
   (ii) Count for the four categories of 'JobSatisfaction' ordered by frequency
3. We want now set a Chi-square test for independence. The null hypothesis H_0 is that there is no significant relationship between 'Attrition' and 'JobSatisfaction'. The alternative hypothesis H_A is that there is significant relationship between 'Attrition' and 'JobSatisfaction'.
   (i) Construct a contingency table using the 'crosstab' function from pandas
   (ii) Calculate the Chi-square statistic
   (iii) Compute the p value
   (iv) Conclude
4. Same using the function `chi2_contingency` of the scipy `library`