Chi-squared Tests for Independence

We recently learnt about the Pearson correlation test for two continuous variables. In some cases, our data may contain categorical variables - to test for the relationship between two categorical variables. The chi-square tests how one categorical variable influences the other categorical variable using the distribution of the frequencies within each group.

We will use the `cars` dataset. Assuming we want to test the relationship between `fuel-type` and `aspiration`. These are two categorical variables and do not have any continuous points to hold them to. It is either the car is a standard car with diesel fuel, a standard car with gas fuel, a turbo car with diesel fuel, or a turbo car with gas fuel. We will like to know if the fuel type is associated (or related) to the aspiration of the car. Before we do let's go through some important points:

1. The Chi-square tests a null hypothesis that the variables are independent. The test compares the observed data to the values that the model expects if the data was distributed in different categories by chance. Anytime the observed data doesn't fit within the model of the expected values, the probability that the variables are dependent becomes stronger, thus proving the null hypothesis incorrect.

2. The Chi-square does not tell you the type of relationship that exists between both variables only that a relationship exists.

Now let us test if there is an association between fuel-type and aspiration.

First, we will find the observed values of cars in each category. This can be done by using the `crosstab` in the pandas library.

	Standard	**Turbo**	**Total**
diesel	7	13	20
gas	161	24	184
Total	168	37	205

We will then calculate the expected values, i.e. the values that we will expect if each car fell into the categories randomly. We do that by using the following formula:

$$\frac{RowTotal * ColumnTotal}{GrandTotal}$$

Using the formula above, expected values for each group will be as follows:

- Standard car with diesel fuel = `(20 * 168)/205`
- Standard car with gas fuel = `(185 * 168)/205`
- Turbo car with diesel fuel = `(20 * 37)/205` and
- Turbo car with gas fuel = `(185 * 37)/205`

We can also look at the expected values in a cross tabular form:

	Standard	Turbo	Total
diesel	16.39	3.61	20
gas	151.61	33.39	184
Total	168	37	205

We can see what the model expects and we can also see that the sub-totals and totals are equal to that of the observed values. We will now perform the Chi-square test. The formula is as follows:

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

This will produce a Chi-square `χ2` value, in this case is `29.6`. We will use the chi-square table and find the corresponding p-value which is done by finding the degree of freedom `(row-1)*(column-1) = 1` and then the corresponding p-value to 29.6, using the chi-square table, we will see that the p-value is very close to 0. This means that there is an **association** between fuel-type and aspiration.

You can do this in python using the `scipy.stats` package, first we create a cross tab: `cont_table = pd.crosstab(df['fuel-type'], df['aspiration'])` then we use the `chi2_contingency` in the `scipy.stats` library: `scipy.stats.chi2_contingency(cont_table, correction = True)`

This will print out the chi-square statistic, the p-value, the degree of freedom, and the expected values.

Author(s)

Aije Egwaikhide

Changelog

Date	Version	Changed by	Change Description
2020-11-24	1.0	Aije	Created the initial version of Chi-square reading document