**Chi-square test of Independence**

A **chi-square (Χ2) test of independence,** also known as a chi-square test of association, is a [nonparametric](https://www.scribbr.com/statistics/statistical-tests/#nonparametric) [hypothesis test](https://www.scribbr.com/statistics/hypothesis-testing/). You can use it to test whether two [categorical variables](https://www.scribbr.com/methodology/types-of-variables/#quantitative-vs-categorical) are related to each other. We can say, If two variables are related, the probability of one variable having a certain value is dependent on the value of the other variable.

NOTE:

1. We cannot use ‘Pearson’s Correlation Coeff. “r”, to find out relationship between two categorical variables because they areused to measure how strong is the relationship **between two QUANTITATIVE variables or two numerical variables or, the Pearson Correlation evaluates whether there is statistical evidence for a linear relationship among the same pairs of variables in the population, represented by a population correlation coefficient, ρ (“rho”).**
2. Pearson’s Correlation Coeff. ‘r’ only reveals associations among continuous variables. But, be careful as the categorical variables may have been encoded use LabelEncoding to appear as Numerical Variable. BUT they should still be treated as Categoric al Variables only and hence a **chi-square (Χ2) test of independence should only be used.**
3. We should not look to find or establish relationship between a Quantitative and a Categorical Variable or vice versa, because it makes no sense.

### **Contingency tables**

When you want to perform a chi-square test of independence, the best way to organize your data is a type of **frequency distribution table** called a **contingency table**.

A contingency table, also known as a cross tabulation or crosstab, shows the number of observations in each combination of groups. It also usually includes row and column totals.

IMPLEMENTATION IN JUPYTER ON PANDA SERIES or ARRAYS

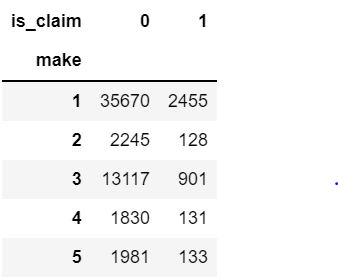
* **pandas.crosstab() function**

This method is used to compute a simple cross-tabulation of two (or more) factors. By default, computes a frequency table (SAME AS CONTINGENCY TABLE) of the factors.

EXAMPLE:

crosstab = pd.crosstab(df["make"], df["is\_claim"])

where, say, df is the dataframe with columns ‘make’ and ‘is\_claim’, both of which are categorical variables.

crosstab -------------------------------------🡪 

**Example:** Contingency table

Six months after the intervention, the city looks at the outcomes for the 300 households (only four households are shown here):

| **Household address** | **Intervention** | **Outcome** |
| --- | --- | --- |
| 25 Elm Street | Flyer | Recycles |
| 100 Cedar Street | Control | Recycles |
| 3 Maple Street | Control | Does not recycle |
| 123 Oak Street | Phone call | Recycles |
| … | … | … |

They reorganize the data into a **contingency table**:

| **Intervention** | **Recycles** | **Does not recycle** | **Row totals** |
| --- | --- | --- | --- |
| Flyer (pamphlet) | 89 | 9 | 98 |
| Phone call | 84 | 8 | 92 |
| Control | 86 | 24 | 110 |
| Column totals | 259 | 41 | N = 300 |

Like all hypothesis tests, the chi-square test of independence evaluates a null and alternative hypothesis. The hypotheses are two competing answers to the question “Are variable 1 and variable 2 related?”

* [Null hypothesis](https://www.scribbr.com/statistics/null-and-alternative-hypotheses/#definition) (*H*0): Variable 1 and variable 2 are **not related** in the population;
* [Alternative hypothesis](https://www.scribbr.com/statistics/null-and-alternative-hypotheses/#alternative) (*H*a): Variable 1 and  variable 2 are **related**in the population.

STEPS TO CARRY OUT CHI-SQUARE TEST:

* 1. Create a table
  2. Create the Contingency table
  3. Calculate value of Pearson’s chi-square (Χ2) value

## How to calculate the test statistic (formula)

Pearson’s chi-square (Χ2) is the [test statistic](https://www.scribbr.com/statistics/test-statistic/) for the chi-square test of independence:

\begin{equation*}X^2 = \sum {\frac {(O-E)^2}{E}}$\end{equation*}

Where

* Χ2 is the chi-square test statistic
* Σ is the summation operator (it means “take the sum of”)
* O is the observed frequency
* E is the expected frequency

### Find the critical chi-square value

### You can find the critical value in a [chi-square critical value table](https://www.scribbr.com/statistics/chi-square-distribution-table/) or using statistical software. You need to known two numbers to find the critical value:

### **The**[**degrees of freedom**](https://www.scribbr.com/statistics/degrees-of-freedom/)**(df):** For a chi-square test of independence, the df is (number of variable 1 groups − 1) \* (number of variable 2 groups − 1).

### [**Significance level (α):**](https://www.scribbr.com/statistics/statistical-significance/#significance-level) By convention, the significance level is usually .05.

Example: For a test of significance at α = .05 and df = 2, the Critical value is 5.99.



The above image shows the first 20 rows of the Chi-Square distribution table, with the degrees of freedom along the left side of the table and the alpha levels (**almost always chosen value is p = 0.05**) along the top of the table.

* 1. Compare the chi-square value to the critical value

Is the test statistic(Χ2) big enough to reject the null hypothesis? Compare it to the critical value

* if it (test statistic, Χ2) is greater, then reject Null Hypothesis(Ho : There is no relation between the two variables) . Thus, conclude that the two variables are related.
* if it (test statistic, Χ2) is smaller, then accept Null Hypothesis(Ho : There is no relation between the two variables) . Thus, conclude that the two variables are NOT related.

NOTE: When you run the hypothesis test, the test will give you a value for p. Compare that value to your chosen alpha level. For example, let’s say you chose an alpha level of 5% (0.05). If the results from the test give you:

* **A small p**(≤ 0.05), [reject the null hypothesis](https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/support-or-reject-null-hypothesis/). This is strong evidence that the null hypothesis is invalid. Therefore, you reject the Null and accept the Alternate Hypothesis H1: that there is relation between the two variables.
* **A large p**(> 0.05) means the [alternate hypothesis](https://www.statisticshowto.com/what-is-an-alternate-hypothesis/) is weak, so you do not reject the null.