

# Data analysis



Chapter4 :

*Correspondence Analysis  
(C.A.) TD*

Dr.Allaoui

08/11/2024

# Table of contents

- I - Chapter4** **3**
- 1. Correspondence Analysis (CA) ..... 3
  - 1.1. 1. Data Preparation ..... 3
  - 1.2. Inertia of the Cloud of Points ..... 9
  - 1.3. Correspondence Analysis (CA) and Links with PCA ..... 11
  - 1.4. Annex ..... 11

# I Chapter4

## 1. Correspondence Analysis (CA)

### Definition

**Correspondence Analysis (CA)**, also called *simple correspondence analysis*, is an *exploratory method* used to analyze *contingency tables*. Its primary objective is to analyze the *relationship between two qualitative variables* by representing categories as points in a 2 or 3 dimensional space graph. So **CA** can be seen as a *special type of Principal Component Analysis (PCA)*, discussed in Chapter 2, using a  $\chi$ -*squared metric* that relies on the *column profiles* of the table.

When there are *more than two* qualitative variables, **Multiple Correspondence Analysis (M.C.A.)**, discussed in Chapter 4, is employed.

### 1.1. 1. Data Preparation

In this type of **Factor Analysis**, we study the *relationships between two variables*,

Consider  $n$  individuals and *two categorical variables*  $V_1$  and  $V_2$ . Let  $x_{pq}$  denote the modality of variable  $V_q$  for individual  $p$ , where  $p = 1, \dots, n$  and  $q = 1, 2$ . These data are represented in a matrix form as:

$$\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{bmatrix}.$$

The modalities of a categorical variable are the different values that this variable can take.

We will denote the  $I$  modalities of the first variable as  $m_{11}, \dots, m_{1i}, \dots, m_{1I}$ , and the  $J$  modalities of the second variable as  $m_{21}, \dots, m_{2j}, \dots, m_{2J}$ .

The goal is to determine if there is an association or relationship between the two variables  $V_1$  and  $V_2$ .

### Deux variables

	<b>V<sub>1</sub></b>	<b>V<sub>2</sub></b>
<b>1</b>	X <sub>11</sub>	X <sub>12</sub>
<b>2</b>	X <sub>21</sub>	X <sub>22</sub>
<b>3</b>	X <sub>31</sub>	X <sub>32</sub>
⋮		
<b>p</b>	X <sub>p1</sub>	X <sub>p2</sub>
⋮		
<b>n</b>	X <sub>n1</sub>	X <sub>n2</sub>

**Q** *Definition: Make a contingency table*

is a type of table used in statistics to display the frequency distribution of variables. It is particularly useful for examining the relationship between two or more categorical variables. Here,  $n_{ij}$  represents the number of individuals who simultaneously possess the modality  $m_{1i}$  of  $V_1$  and the modality  $m_{2j}$  of  $V_2$ . It is easy to see that:

$$n = \sum_{i=1}^I \sum_{j=1}^J n_{ij}$$

The contingency table is as follows:

	<i>Modalités de V<sub>2</sub></i>					
	<b>m<sub>21</sub></b>	<b>m<sub>22</sub></b>		<b>m<sub>2j</sub></b>		<b>m<sub>2J</sub></b>
<b>m<sub>11</sub></b>	n <sub>11</sub>	n <sub>12</sub>	...	n <sub>1j</sub>	...	n <sub>1J</sub>
<b>m<sub>12</sub></b>	n <sub>21</sub>	n <sub>22</sub>	...	n <sub>2j</sub>		n <sub>2J</sub>
	⋮	⋮				⋮
<b>m<sub>1i</sub></b>	n <sub>i1</sub>	n <sub>i2</sub>	...	n <sub>ij</sub>		n <sub>iJ</sub>
	⋮	⋮				⋮
<b>m<sub>1I</sub></b>	n <sub>I1</sub>	n <sub>I2</sub>	...	n <sub>Ij</sub>	...	n <sub>IJ</sub>

**Q** *Example*

For this example, we have 30 items (office furniture) and two variables: the variable Type (type of furniture) and the variable Color (furniture color). We then construct a contingency table:

	gray	brown	black	TOTAL	%
cabinet	1	3	5	9	0.3
desk	2	6	3	11	0.367
chair	5	4	1	10	0.333
TOTAL	8	13	9	30	1
%	0.267	0.433	0.3	1	

**Q** *Definition: Calculation of Relative Frequencies ( Probabilities)*

---

Next, compute the relative frequencies for each cell in the table. This is done as :

$$f_{ij} = \frac{n_{ij}}{n};$$

where  $f_{ij}$  represents **the joint probability** of observing both modality  $m_{1i}$  of variable  $V_1$  and modality  $m_{2j}$  of variable  $V_2$ . In this framework, the column margin (and, similarly, the row margin) corresponds to the column profile (or, respectively, the row profile).

The probability table derived from our contingency table gives added interpretive value to the data. For the current example, we obtain :

The row marginal  $f_{i\cdot}$  is the sum of all relative frequencies in row  $i$ , and the column marginal  $f_{\cdot j}$  is the sum of all relative frequencies in column  $j$ . They are defined as:

$$f_{i\cdot} = \sum_j f_{ij}$$

$$f_{\cdot j} = \sum_i f_{ij}$$

	gris	marron	noir	$C_m$
armoire	0,033	0,100	0,167	0.300
bureau	0,067	0,200	0,100	0.367
chaise	0,167	0,133	0,033	0.333
$L_m$	0,267	0,433	0,300	1

**Note**

---

If the two qualitative variables are independent, performing Correspondence Analysis is not meaningful.⇒ Start by conducting a Chi-square ( $\chi^2$ ) test.

### Standardization and Calculation of Chi-Square Distances:

Next, calculate the standardized residuals or **chi-square distances**, which *measure how much the observed frequencies deviate from what would be expected if the two variables were independent*. The chi-square distance is computed as:

$$\chi^2 = n \sum_{i=1}^I \sum_{j=1}^J \frac{(f_{ij} - f_{i \cdot} f_{\cdot j})^2}{f_{i \cdot} f_{\cdot j}}$$

Here,  $\chi^2$  *represents the significance of the relationship between the variables. It measures the difference between the observed frequencies and the expected (theoretical) frequencies.*

In the above example, the calculations yield:

$$\chi^2 = n \sum_{i=1}^I \sum_{j=1}^J \frac{1}{f_{i \cdot} f_{\cdot j}} (f_{ij} - f_{i \cdot} f_{\cdot j})^2 = 75.35.$$

### Understanding Row and Column Profiles in Contingency Tables

In data analysis, contingency tables allow us to observe the relationship between two categorical variables. To better understand the distribution patterns within these tables, we can use **row profiles** and **column profiles**. These profiles help us analyze the proportional distribution of frequencies within each row or column of the table.

- **Row Profiles** : for a **contingency table** with rows ( $i = 1, \dots, I$ ) and columns ( $j = 1, \dots, J$ ), the row profile  $p_{ij}$  for each cell is calculated as follows:

$$p_{ij} = \frac{f_{ij}}{f_{i \cdot}}$$

where:

- $f_{ij}$  is the observed frequency in row  $i$  and column  $j$ ,
- $f_{i \cdot}$  is the total frequency for row  $i$ , calculated as  $f_{i \cdot} = \sum_{j=1}^J f_{ij}$ .
- **Interpret the row Profiles**: show the **proportion of each cell in a row** relative to **the total frequency of that row**. This allows us to compare how the values in a row are distributed across the columns.

	gris	marron	noir	TOTAL	$C_m$
armoire	1	3	5	9	0.300
$L_1$	0,111	0,333	0,556	1	
bureau	2	6	3	11	0.367
$L_2$	0,182	0,545	0,273	1	
chaise	5	4	1	10	0.333
$L_3$	0,500	0,400	0,100	1	
TOTAL	5	10	9	30	1
$L_m$	0,167	0,333	0,300	1	

We can express  $L_m$  as:

$$L_m = \frac{n_1}{n}L_1 + \frac{n_2}{n}L_2 + \frac{n_3}{n}L_3$$

$$L_m = f_1.L_1 + f_2.L_2 + f_3.L_3$$

This equation represents **the weighted average (mean) of the row profiles**  $L_1$ ,  $L_2$ , and  $L_3$ , where the weights are given by the proportions  $f_1$ ,  $f_2$ , and  $f_3$ .

	$m2_1$	$m2_2$	$m2_3$	$C_m$
$L_1$	$f_{11}/f_1$	$f_{12}/f_1$	$f_{13}/f_1$	$f_1$
$L_2$	$f_{21}/f_2$	$f_{22}/f_2$	$f_{23}/f_2$	$f_2$
$L_3$	$f_{31}/f_3$	$f_{32}/f_3$	$f_{33}/f_3$	$f_3$
$L_m$	$f_{.1}$	$f_{.2}$	$f_{.3}$	1

- **Construction of Column Profiles** : In correspondence analysis or contingency table analysis, **column profiles** represent the distribution of each column variable across different row categories. Constructing these profiles allows us to understand how each column category is associated with each row category, typically in a percentage format.

**Calculate Column Profiles:** For each cell  $f_{ij}$  in the frequency table, calculate the proportion of  $f_{ij}$  relative to the total of its column  $f_{.j}$ :

$$q_{ij} = \frac{f_{ij}}{f_{.j}}$$

where  $p_{ij}$  represents the column profile for row  $i$  in column  $j$ .

- **Interpret the Column Profiles:** The value  $p_{ij}$  shows **the proportion** of column  $j$  that falls into row  $i$ . **This helps in identifying patterns or associations between row and column categories.**
- **Compute Column Totals:** For each column in the contingency table, sum all the observed values in that column. Denote these totals as  $f_{.j}$  for each column  $j$ .

	gris	$C_1$	marron	$C_2$	noir	$C_3$	TOTAL	$C_m$
armoire	1	0,125	3	0,231	5	0,556	9	0,300
bureau	2	0,250	6	0,462	3	0,333	11	0,367
chaise	5	0,625	4	0,308	1	0,111	10	0,333
TOTAL	8	1	13	1	9	1	30	1
$L_m$	0,267		0,433		0,300		1	

The column that includes the total distribution of all categories, called the **mean column profile**, is denoted by  $C_m$ .

**Calculating the Mean Column Profile :**

For each row  $i$ , we define the mean column profile  $C_m$  as a weighted average of  $C_1$ ,  $C_2$ , and  $C_3$ , where each column profile is weighted by the relative proportions  $f_{.1}$ ,  $f_{.2}$ , and  $f_{.3}$ , given by:

$$C_m = f_{.1}C_1 + f_{.2}C_2 + f_{.3}C_3.$$

**Interpretation of the Mean Column Profile**

The mean column profile  $C_m$  provides **an average representation of the distribution across all categories in the table**, with each category's profile weighted by its relative size in the data. This approach offers a balanced view, capturing the overall structure of the variable "Color" while taking into account the varying sizes of each category.

## ⚙️ Method

The following table, based on the structure of the example discussed, summarizes **the notations used above to analyze the relationships between qualitative variables**.

	$m2_1$	$C_1$ %	$m2_2$	$C_2$ %	$m2_3$	$C_3$ %	TOTAL	$C_m$ %
$m1_1$	$f_{11}$	$f_{11}/f_{.1}$	$f_{12}$	$f_{12}/f_{.2}$	$f_{13}$	$f_{13}/f_{.3}$	$n_1$	$f_{.1}$
$L_1$ %	$f_{11}/f_{.1}$	$f_{.1}.f_{.1}$	$f_{12}/f_{.1}$	$f_{.1}.f_{.2}$	$f_{13}/f_{.1}$	$f_{.1}.f_{.3}$	1	
$m1_2$	$f_{21}$	$f_{21}/f_{.1}$	$f_{22}$	$f_{22}/f_{.2}$	$f_{23}$	$f_{23}/f_{.3}$	$n_2$	$f_{.2}$
$L_2$ %	$f_{21}/f_{.2}$	$f_{.2}.f_{.1}$	$f_{22}/f_{.2}$	$f_{.2}.f_{.2}$	$f_{23}/f_{.2}$	$f_{.2}.f_{.3}$	1	
$m1_3$	$f_{31}$	$f_{31}/f_{.1}$	$f_{32}$	$f_{32}/f_{.2}$	$f_{33}$	$f_{33}/f_{.3}$	$n_3$	$f_{.3}$
$L_3$ %	$f_{31}/f_{.3}$	$f_{.3}.f_{.1}$	$f_{32}/f_{.3}$	$f_{.3}.f_{.2}$	$f_{33}/f_{.3}$	$f_{.3}.f_{.3}$	1	
TOTAL	$n_1$	1	$n_2$	1	$n_3$	1	$n$	1
$L_m$ %	$f_{.1}$		$f_{.2}$		$f_{.3}$			1

## Matrix Notations

We will use matrix notations for the calculations developed below. Let  $P = (f_{ij})$  denote the matrix of probabilities,

$\mathbb{1}_I$  be the unit vector in  $\mathbb{R}^I$ :

$$\mathbb{1}_I = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix},$$

and  $\mathbb{1}_J$  the unit vector in  $\mathbb{R}^J$ :

$$\mathbb{1}_J = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

The **row mean profile** is given by  $L_m = P^T \mathbb{1}_I$ , and **the column mean profile** is  $C_m = P \mathbb{1}_J$ .

Recall that:

$$L_m = \begin{pmatrix} f_{.1} \\ f_{.2} \\ \vdots \\ f_{.J} \end{pmatrix}, C_m = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_J \end{pmatrix}.$$

Let  $L = \left( \frac{f_{ij}}{f_{.i}} \right)$  represent the row profile matrix. For our example, we have:

$$L = \begin{bmatrix} f_{11}/f_{.1} & f_{12}/f_{.1} & f_{13}/f_{.1} \\ f_{21}/f_{.2} & f_{22}/f_{.2} & f_{23}/f_{.2} \\ f_{31}/f_{.3} & f_{32}/f_{.3} & f_{33}/f_{.3} \\ f_{41}/f_{.4} & f_{42}/f_{.4} & f_{43}/f_{.4} \\ f_{51}/f_{.5} & f_{52}/f_{.5} & f_{53}/f_{.5} \end{bmatrix}.$$



Let  $\mathbf{D}_I = \text{diag}(f_i)$  and  $\mathbf{D}_J = \text{diag}(f_{\cdot j})$  denote the diagonal matrices whose diagonal elements are the components of  $\mathbf{C}_m$  and  $\mathbf{L}_m$ , respectively. We verify that:

$$\mathbf{L} = \mathbf{D}_I^{-1} \mathbf{P}, \quad \mathbf{C} = \mathbf{D}_J^{-1} \mathbf{P}^\top.$$

## 1.2. Inertia of the Cloud of Points

### Centroid $G$ of Cloud

Each point in the cloud  $\mathfrak{N}_I$ , indexed by  $i$ , is assigned a weight  $p_i$ . The centroid  $G$ , located in  $\mathbb{R}^J$ , has its  $j$ -th coordinate calculated as:

$$G_j = \sum_{i \in I} p_i \frac{f_{ij}}{p_i} = \sum_{i \in I} f_{ij} = f_{\cdot j}.$$

$$G_i = \sum_{j \in J} p_{\cdot j} \frac{f_{ij}}{p_{\cdot j}} = \sum_{j \in J} f_{ij} = f_i.$$

Thus, the coordinates of the centroid  $G$  are equivalent to the marginal weights of  $\mathfrak{N}_I$ .

### Chi-Square Distance in Correspondence Analysis (CA)

In CA, **the similarity between row or column profiles is measured using the  $\chi^2$ -distance**. It calculates the weighted sum of squared differences between two profiles.

#### The distance between rows

A distance metric is defined within the  $\mathbb{R}^J$  space to account for the nature of the points. Specifically :

- **The distance between a row profile  $\mathbf{L}_r$  and the row profile  $\mathbf{L}_s$  becomes:**

$$d_{\chi^2}^2(L_r, L_s) = \sum_j \frac{1}{f_{\cdot j}} \left( \frac{f_{rj}}{f_{\cdot r}} - \frac{f_{sj}}{f_{\cdot s}} \right)^2.$$

or

$$d_{\chi^2}^2(L_r, L_s) = \sum_{j=1}^J \frac{(f_{rj} - f_{sj})^2}{p_j} = \frac{(f_{1r} - f_{1s})^2}{p_1} + \dots + \frac{(f_{jr} - f_{js})^2}{p_j} + \dots + \frac{(f_{jr} - f_{js})^2}{p_j}.$$

- **The distance the a row profile  $\mathbf{L}_r$  and mean rows profile:**

$$d_{\chi^2}^2(L_r, G_j) = \sum_j \frac{1}{f_{\cdot j}} \left( \frac{f_{rj}}{f_{\cdot r}} - f_{\cdot j} \right)^2.$$

#### The distance between Columns

$$d_{\chi^2}^2(C_r, C_s) = \sum_i \frac{1}{f_i} \left( \frac{f_{ir}}{f_{\cdot r}} - \frac{f_{is}}{f_{\cdot s}} \right)^2.$$

- **The distance the a Column profile  $\mathbf{L}_r$  and mean Columns profile:**

$$d_{\chi^2}^2(C_r, G_j) = \sum_i \frac{1}{f_i} \left( \frac{f_{ir}}{f_{\cdot r}} - f_i \right)^2.$$

🔗 *Note: Important property of this metric:*

---

**Merging Similar Categories:** If two columns are proportional (very similar), merging them does not change the distance between row profiles. This means similar categories can be combined without affecting the results. As example:

If the categories "skilled workers" and "unskilled workers" have similar profiles, they can be grouped together without changing the analysis results.

We can see, in particular, that if the two variables are independent, then  $f_{ij} = f_{i.}f_{.j}$  for all  $i = 1, \dots, I$  and  $j = 1, \dots, J$ .

Thus:

$$d_{\chi^2}^2(\mathbf{L}_i, \mathbf{L}_m)^2 = 0$$

This implies that **all points in the cloud coincide with the mean row profile  $\mathbf{L}_m$ .**

### *Expression of the khi sequered Distance and Inertia*

As explained in the course on PCA, the dispersion of a cloud of points noted  $\mathfrak{N}_J$  quantifies the amount of information it contains. Each point  $L_i$  has a weight, reflecting its importance. The dispersion is measured by calculating the **inertia of the cloud of  $L_r$  relative to  $L_s$** , using the formula:

$$\mathcal{I} = \sum_i f_i \cdot d_{\chi^2}^2(L_r, L_s).$$

*The total inertia of the cloud of points formed by the row profiles.*

The metric associated with the distance  $d_{\chi^2}^2$ , as defined in the previous section, is referred to as the "metric of the  $\chi^2$  Test. This metric establishes a relationship between the calculation of the  $\chi^2$  test (used to assess the independence of two variables) and the **inertia** of the clouds  $\mathfrak{N}_I$  and  $\mathfrak{N}_J$  with respect to the centroid  $G$ , calculated using the  $\chi^2$  metric.

We have:

$$\mathcal{J} = \sum_{i=1, I} \sum_{j=1, J} \frac{1}{f_i f_j} (f_{ij} - f_i f_j)^2.$$

Where  $\mathcal{J}$  represents **the total inertia of the cloud of points formed by the row profiles.**

We can deduce that :

$$\chi^2 = n\mathcal{J}.$$

**Interpretation:** In statistical terms, inertia can be **interpreted as the total variance** or spread of the data when projected onto a lower-dimensional space. It helps measure how well the data is represented in that space.

In the **case of Correspondence Analysis**, inertia **helps quantify the amount of information** (or variability) captured by the axes in the reduced-dimensional representation of the contingency table. **Higher inertia corresponds to more variation captured**, and lower inertia indicates less variation or poorer representation.

### 1.3. Correspondence Analysis (CA) and Links with PCA

The *relationship* between two variables can be better *understood* by *analyzing the distinctions between their profiles* (rows and columns of a contingency table). PCA is a powerful tool for this purpose, providing graphical representations of these profiles and highlighting their similarities and differences. This type of analysis involves complex mathematical computations, particularly the diagonalization of non-symmetric matrices, which cannot be done manually. Modern computational tools simplify this process, enabling us to focus on interpreting the visualizations.

By applying PCA to both row and column profiles and combining the results into a single scatter plot, we gain a clear and comprehensive understanding of the associations between the two variables. This theoretical foundation serves as a critical step toward practical applications of PCA in analyzing multivariate data.

#### *Method: The Methodology*

---

##### - *PCA on Row Profiles:*

We first perform a PCA on the row profiles of the contingency table. In this case, the "individuals" in the PCA are the rows of the table, which represent the categories or levels of variable X.

*The goal* is to represent these *row profiles graphically*, i.e., to visualize the modalities (values) of X. In this specific PCA, we focus on the rows, not on the variables themselves. If only *two dimensions* are retained (as in our case), we get a *single graph* if more dimensions (in case of *MCA* in the next chapter) are preserved, *multiple graphs* are produced.

##### - *PCA on Column Profiles:*

Next, we perform a PCA on the column profiles of the contingency table. Here, the "individuals" are the columns, which correspond to the categories or levels of variable Y.

A graphical representation is created for these column profiles, thus visualizing the modalities of Y.

##### - *Linking the Two Analyses:*

It is important to note that these two PCA analyses are closely related, as they are both derived from the same contingency table. As a result, it is valid to combine the two graphical representations. By doing so, we obtain a scatter plot (or a series of scatter plots if more than two dimensions are kept), which represents both the modalities of X and Y simultaneously.

### 1.4. Annex

#### *Reminder: Independence Situation*

---

Consider a contingency table where the rows correspond to the modalities of variable  $V_1$  and the columns correspond to the modalities of variable  $V_2$ . Each cell  $f_{ij}$  *represents the number of observations where  $V_1$  takes modality  $m_{1i}$  and  $V_2$  takes modality  $m_{2j}$ .*

##### - *Conditional and Marginal Probabilities :*

*Conditional Probability:* the probability of observing modality  $m_{2j}$  of  $V_2$  given that  $V_1$  takes modality  $m_{1i}$  is given by:

$$P(m_{2j} | m_{1i}) = \frac{f_{ij}}{f_{i.}}$$

where  $f_{i.}$  is the total of row  $i$ , i.e., the marginal frequency of modality  $m_{1i}$ .

- **Marginal Probability:** The probability of observing modality  $m_{2j}$  of  $V_2$  without any condition on  $V_1$  is given by:

$$P(m_{2j}) = \frac{f_{.j}}{n},$$

where  $f_{.j}$  is the total of column  $j$ , i.e., the marginal frequency of modality  $m_{2j}$ .

- **Independence of Variables**

The variables  $V_1$  and  $V_2$  are said to be independent if, for all  $i$  and  $j$ , the conditional probability of  $m_{2j}$  given  $m_{1i}$  equals the marginal probability of  $m_{2j}$ :

$$P(m_{2j} | m_{1i}) = P(m_{2j}).$$

This implies that:

$$f_{ij} = \frac{f_{i.} \times f_{.j}}{n}.$$

In other words, **the observed frequencies  $f_{ij}$  must equal the expected frequencies under the assumption of independence.** The observed contingency table is then equal to the table of expected frequencies (or frequencies under independence).

### **Warning: Table in the Independence Situation**

Under the assumption of independence, the Joint Probability Table equals the Marginal Probability Table.

	$m_{2_1}$	$m_{2_2}$	...	$m_{2_j}$	...	$m_{2_J}$	$C_m$
$m_{1_1}$	$f_{11}$	$f_{12}$	...	$f_{1j}$	...	$f_{1J}$	$f_{1.}$
$m_{1_2}$	$f_{21}$	$f_{22}$	...	$f_{2j}$	...	$f_{2J}$	$f_{2.}$
	$\vdots$	$\vdots$				$\vdots$	$\vdots$
$m_{1_i}$	$f_{i1}$	$f_{i2}$	...	$f_{ij}$	...	$f_{iJ}$	$f_{i.}$
	$\vdots$	$\vdots$				$\vdots$	$\vdots$
$m_{1_1}$	$f_{11}$	$f_{12}$	...	$f_{1j}$	...	$f_{1J}$	$f_{1.}$
$L_m$	$f_{.1}$	$f_{.2}$	...	$f_{.j}$	...	$f_{.J}$	1

=

	$m_{2_1}$	$m_{2_2}$	...	$m_{2_j}$	...	$m_{2_J}$	$C_m$
$m_{1_1}$	$f_{1.} \cdot f_{.1}$	$f_{1.} \cdot f_{.2}$	...	$f_{1.} \cdot f_{.j}$	...	$f_{1.} \cdot f_{.J}$	$f_{1.}$
$m_{1_2}$	$f_{2.} \cdot f_{.1}$	$f_{2.} \cdot f_{.2}$	...	$f_{2.} \cdot f_{.j}$	...	$f_{2.} \cdot f_{.J}$	$f_{2.}$
	$\vdots$	$\vdots$				$\vdots$	$\vdots$
$m_{1_i}$	$f_{i.} \cdot f_{.1}$	$f_{i.} \cdot f_{.2}$	...	$f_{i.} \cdot f_{.j}$	...	$f_{i.} \cdot f_{.J}$	$f_{i.}$
	$\vdots$	$\vdots$				$\vdots$	$\vdots$
$m_{1_1}$	$f_{1.} \cdot f_{.1}$	$f_{1.} \cdot f_{.2}$	...	$f_{1.} \cdot f_{.j}$	...	$f_{1.} \cdot f_{.J}$	$f_{1.}$
$L_m$	$f_{.1}$	$f_{.2}$	...	$f_{.j}$	...	$f_{.J}$	1

### **Conclusion on Independence**

The variables  $V_1$  and  $V_2$  are independent if and only if the observed contingency table is equal to the expected frequency table, i.e.,

$$f_{ij} = \frac{f_{i.} \times f_{.j}}{n}.$$

**In practice**, we compare the observed frequencies  $f_{ij}$  with the expected frequencies to **test this independence**, often using the  **$\chi$ -squared ( $\chi^2$ ) test**.

**$\chi$ -Square Test of Independence :** The  $\chi$ -Square Test of Independence is a statistical test used to **determine whether two categorical variables are independent or dependent**. It is based on comparing the observed frequencies with the expected frequencies under the assumption of independence.

**Hypotheses of the Test**

- **Null Hypothesis  $H_0$ :** The variables  $V_1$  and  $V_2$  are independent.
- **Alternative Hypothesis  $H_1$ :** The variables  $V_1$  and  $V_2$  are dependent.

The  $\chi$ -Square Test of Independence is used to validate or reject the null hypothesis ( $H_0$ ) in favor of the alternative hypothesis ( $H_1$ ).

This test assesses the deviation from independence by comparing the joint probability table with the table of products of the marginal probabilities, or equivalently, the contingency table with the table of theoretical frequencies.

Under the null hypothesis  $H_0$ , the variables  $V_1$  and  $V_2$  are independent. Therefore, for all  $i = 1, \dots, I$  and  $j = 1, \dots, J$ , we should have:

$$f_{ij} = f_{i.} \cdot f_{.j}$$

where  $f_{ij}$  represents the observed frequency,  $f_{i.}$  represents the marginal total of row  $i$ , and  $f_{.j}$  represents the marginal total of column  $j$ .

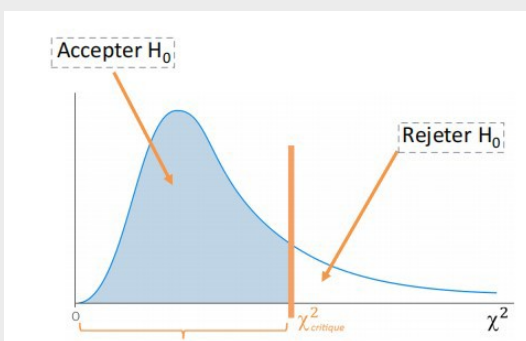
In strict terms, the perfect equality  $f_{ij} = f_{i.} \cdot f_{.j}$  is not required. Instead, we define a significance threshold  $\alpha$ , beyond which the null hypothesis  $H_0$  is rejected.

Typically,  $\alpha = 0.05$  is chosen, but other values of  $\alpha$  can be used depending on the context of the problem.

**Conditions for Applying the Chi-Square Test :**

The Chi-Square test requires the following conditions to be met:

1. The observations used to create the contingency table must be random.
2. The random sample must have a size  $n \geq 30$ .



**Decision Rules for the Chi-Square Test:**

To make a decision in the Chi-Square test, a critical value  $\chi^2_{critical}$  is determined based on the probability distribution used, which depends on the significance level  $\alpha$  and the degrees of freedom  $\nu$ .

The decision rules are as follows:

- If the observed value  $\chi^2 < \chi^2_{critical}$ , then we fail to reject the null hypothesis  $H_0$ .
- If the observed value  $\chi^2 \geq \chi^2_{critical}$ , then we reject the null hypothesis  $H_0$ .