Data analysis

جامعة باجي مختار - عنابة

كلية العلوم الاقتصادية وعلوم التسيير

FACULTY OF ECONOMICS AND MANAGEMENT

Badji Mokhtar – Annaba University

Chapter4 :
Correspondence Analysis
(C.A.)

Dr.Allaoui

08/11/2024

Table of contents

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 χ***-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***,

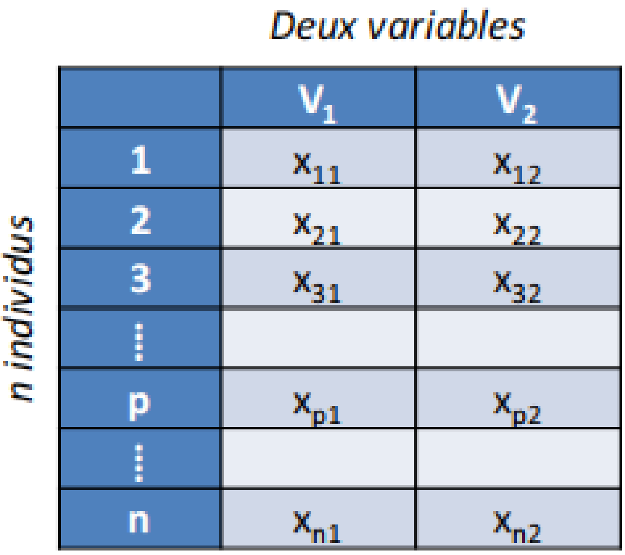
Consider n individuals and ***two categorical variables*** V_1 and V_2. Let x_{pq} denote the modality of variable V_q for individual p, where $p = 1, \ldots, n$ and $q = 1, 2$. These data are represented in a matrix form as:

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

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 $m_{11}, \ldots, m_{1i}, \ldots, m_{1I}$, and the J modalities of the second variable as $m_{21}, \ldots, m_{2j}, \ldots, m_{2J}$.

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



Deux variables

n individus

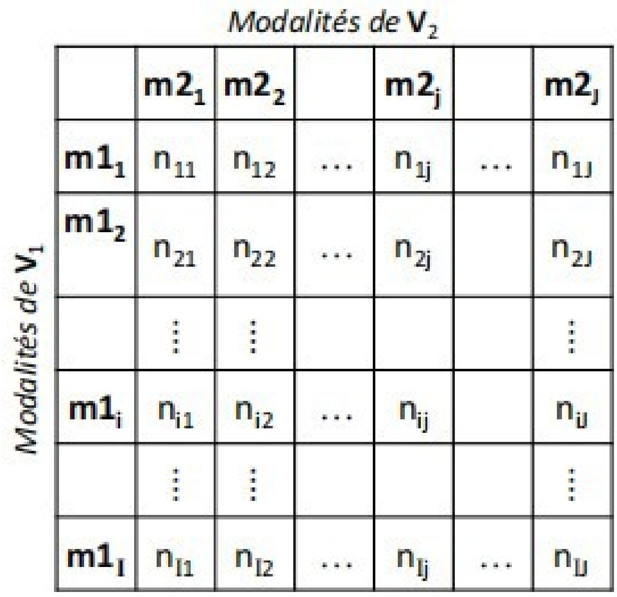
	V_1	V_2
1	x_{11}	x_{12}
2	x_{21}	x_{22}
3	x_{31}	x_{32}
⋮		
p	x_{p1}	x_{p2}
⋮		
n	x_{n1}	x_{n2}

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:



Modalités de V_2

Modalités de V_1

	$m2_1$	$m2_2$		$m2_j$		$m2_J$
$m1_1$	n_{11}	n_{12}	…	n_{1j}	…	n_{1J}
$m1_2$	n_{21}	n_{22}	…	n_{2j}		n_{2J}
	⋮	⋮				⋮
$m1_i$	n_{i1}	n_{i2}	…	n_{ij}		n_{iJ}
	⋮	⋮				⋮
$m1_I$	n_{I1}	n_{I2}	…	n_{Ij}	…	n_{IJ}

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	

Definition: Calculation of Relative Frequencies (Probabilities)

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

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

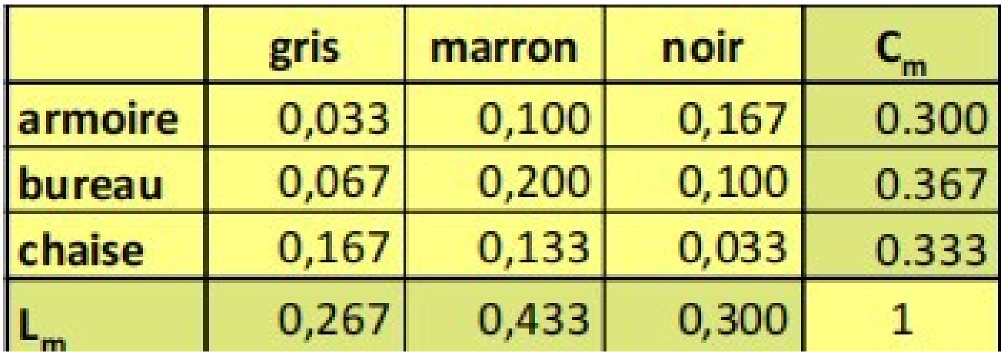
where f_{ij} represents ***the joint probability*** of observing both modality m_{1i} of variable V_1 and modality m_{2j} of variable V_2. In this framework, the column margin (and, similarly, the row margin) corresponds to the column profile (or, respectively, the row profile).

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

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

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

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



	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 (χ^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:

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

Here, χ^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:

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

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 $(i = 1, \ldots, I)$ and columns $(j = 1, \ldots, J)$, the row profile p_{ij} for each cell is calculated as follows:

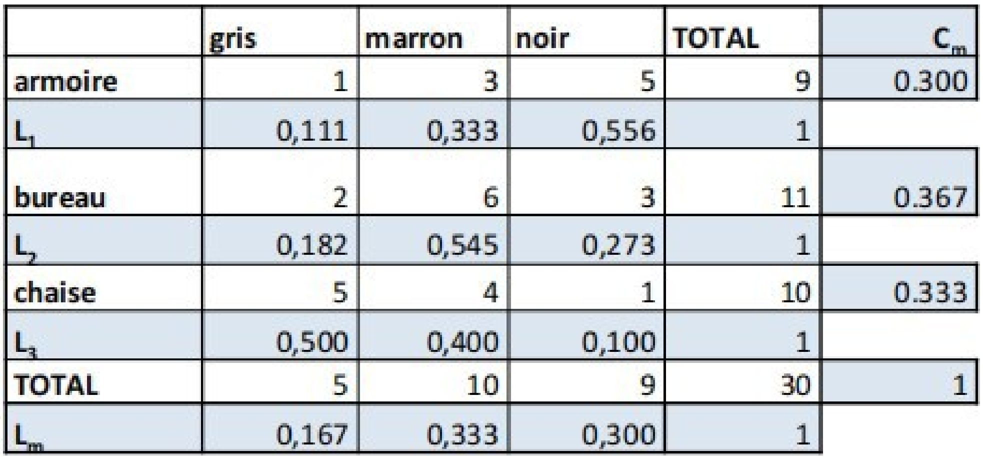
$$p_{ij} = \frac{f_{ij}}{f_{i\cdot}}$$

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:

$$L_m = \frac{n_{1\cdot}}{n}L_1 + \frac{n_{2\cdot}}{n}L_2 + \frac{n_{3\cdot}}{n}L_3$$

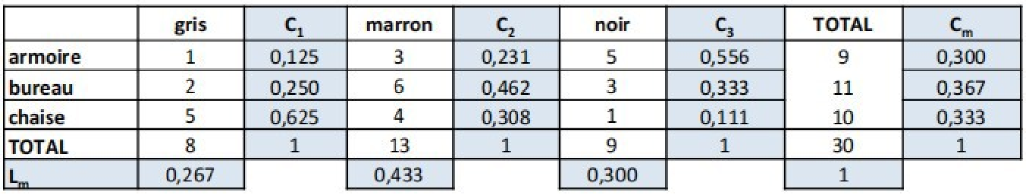
This equation represents ***the weighted average (mean) of the row profiles*** L_1, L_2, and L_3, where the weights are given by the proportions $\frac{n_{1\cdot}}{n}$, $\frac{n_{2\cdot}}{n}$, and $\frac{n_{3\cdot}}{n}$.

- ***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 frequancy table, calculate the proportion of f_{ij} relative to the total of its column $f_{.j}$:

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

  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:

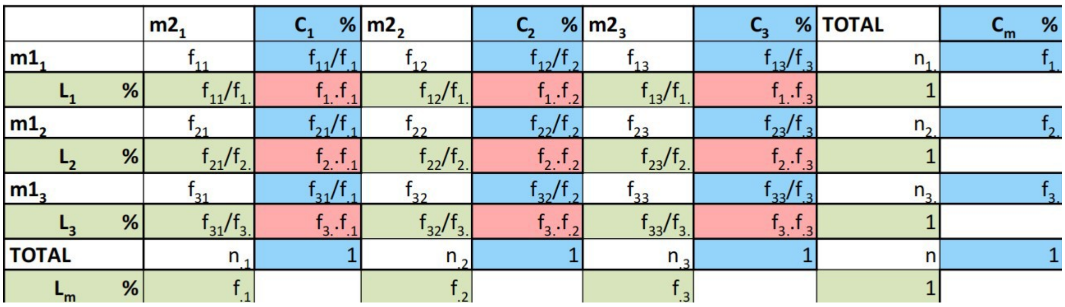
$$C_m = f_{.1}C_1 + f_{.2}C_2 + f_{.3}C_3.$$

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 $P = \left(f_{ij}\right)$ 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 $\mathbf{L}_m = P^\top \mathbb{1}_I$, and ***the column mean profile*** is $\mathbf{C}_m = P \mathbb{1}_J$.

Recall that:

$$\mathbf{L}_m = \begin{pmatrix} f_{\cdot 1} \\ f_{\cdot 2} \\ \vdots \\ f_{\cdot J} \end{pmatrix}, \mathbf{C}_m = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_I \end{pmatrix}.$$

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

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

Let $\mathbf{D}_I = \operatorname{diag}(f_i)$ and $\mathbf{D}_J = \operatorname{diag}(f_{\cdot j})$ 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

Expression of the khi sequered Distance and Inertia

As explained in the course on PCA, ***the dispersion of a cloud of points 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:

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

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 mean row profile $\mathbf{L}_s$ becomes:

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

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

Thus:

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

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

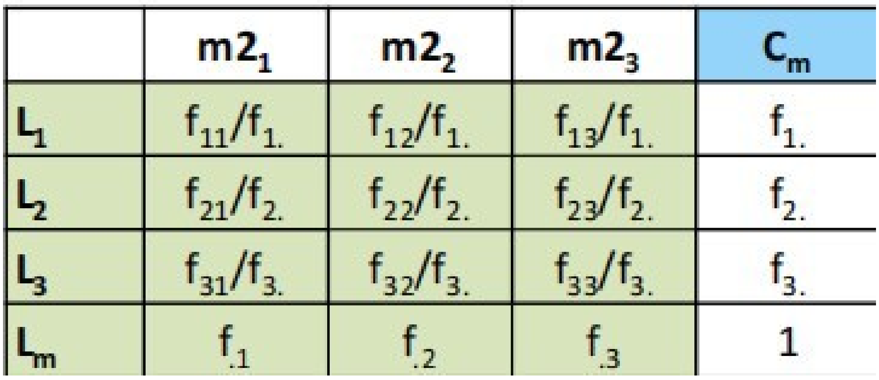
We saw that:

$$\mathcal{J} = \sum_i f_{i\cdot}.d\,(L_i, L_m)^2 = \sum_i f_{i\cdot} \sum_{j=1}^{J} f_{\cdot j} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{sj}}{f_{s\cdot}}\right)^2$$

This can also be written as:

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

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



	m2₁	m2₂	m2₃	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

Matrix notation

Using matrix notation, we get:

$$\mathcal{J} = \mathrm{tr}\left(\mathbf{D}_I^{-1}\left(\mathbf{P} - \mathbf{C}_m\mathbf{L}_m^T\right)\mathbf{D}_J^{-1}\left(\mathbf{P} - \mathbf{C}_m\mathbf{L}_m^T\right)^T\right)$$

1.3. 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:

$$P(m_{2j} \mid m_{1i}) = \frac{f_{ij}}{f_{i\cdot}},$$

where $f_{i\cdot}$ 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:

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

where $f_{\cdot j}$ 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}:

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

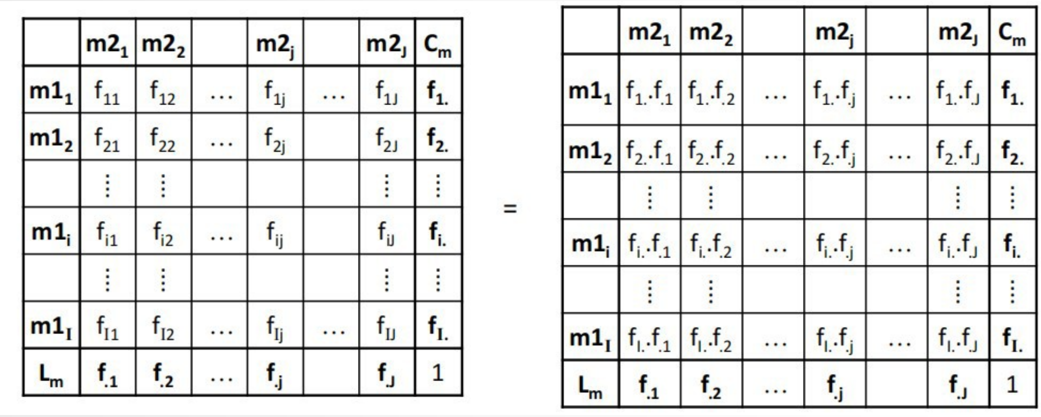
This implies that:

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

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.



	$m2_1$	$m2_2$		$m2_j$		$m2_J$	C_m
$m1_1$	f_{11}	f_{12}	…	f_{1j}	…	f_{1J}	$f_{1.}$
$m1_2$	f_{21}	f_{22}	…	f_{2j}		f_{2J}	$f_{2.}$
	⋮	⋮				⋮	⋮
$m1_i$	f_{i1}	f_{i2}	…	f_{ij}		f_{iJ}	$f_{i.}$
	⋮	⋮				⋮	⋮
$m1_I$	f_{I1}	f_{I2}	…	f_{Ij}	…	f_{IJ}	$f_{I.}$
L_m	$f_{.1}$	$f_{.2}$	…	$f_{.j}$		$f_{.J}$	1

=

	$m2_1$	$m2_2$		$m2_j$		$m2_J$	C_m
$m1_1$	$f_{1.}.f_{.1}$	$f_{1.}.f_{.2}$	…	$f_{1.}.f_{.j}$	…	$f_{1.}.f_{.J}$	$f_{1.}$
$m1_2$	$f_{2.}.f_{.1}$	$f_{2.}.f_{.2}$	…	$f_{2.}.f_{.j}$	…	$f_{2.}.f_{.J}$	$f_{2.}$
	⋮	⋮				⋮	⋮
$m1_i$	$f_{i.}.f_{.1}$	$f_{i.}.f_{.2}$	…	$f_{i.}.f_{.j}$	…	$f_{i.}.f_{.J}$	$f_{i.}$
	⋮	⋮				⋮	⋮
$m1_I$	$f_{I.}.f_{.1}$	$f_{I.}.f_{.2}$	…	$f_{I.}.f_{.j}$	…	$f_{I.}.f_{.J}$	$f_{I.}$
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.,

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

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

χ***-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 $i = 1, \ldots, I$ and $j = 1, \ldots, J$, we should have:

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

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

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

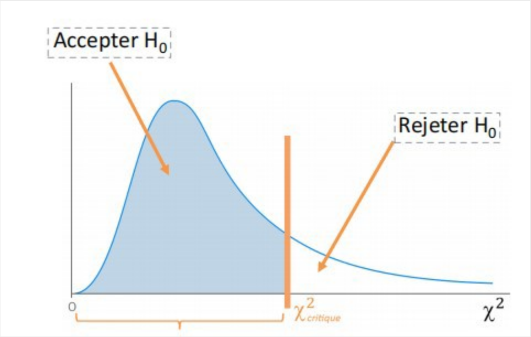
Typically, $\alpha = 0.05$ is chosen, but other values of α can be used depending on the context of the problem.

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 χ^2_{critical} is determined based on the probability distribution used, which depends on the significance level α and the degrees of freedom ν. The decision rules are as follows:

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

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