STATISTICAL LABORATORY

Applied Mathematics for Economics and Management
1st Year/1st Semester
2025/2026

CONTACT

Professor: Elisabete Fernandes
E-mail: efernandes@iseg.ulisboa.pt

https://doity.com.br/estatistica-aplicada-a-nutricao

https://basiccode.com.br/produto/informatica-basica/

PROGRAM

1. Fundamental Concepts of Statistics
2. Exploratory Data Analysis
3. Organizing and Summarizing Data
4. Association and Relationships Between Variables
5. Index Numbers
6. Time Series Analysis

LECTURE 7: CONTINGENCY TABLES AND MEASURES OF ASSOCIATION

CONTINGENCY TABLES: DEFINITION AND EXAMPLE

Definition

- A **contingency table** summarizes the frequency distribution of **two categorical variables**.
- It helps analyze the **association between variables**.

General Example of a Contingency Table

- Let variable **A** have categories $A_1, A_2, \ldots, A_I$ and variable **B** have categories $B1, B_2, \ldots, B_c$.

	B_1	B_2	---	B_c	Totals
A_1	n_{11}	n_{12}	…	n_{1c}	$n_{1.}$
A_2	n_{21}	n_{22}	…	n_{2c}	$n_{2.}$
…	…	…	…	…	…
A_I	n_{I1}	n_{I2}	…	n_{Ic}	$n_{I.}$
Totals	$n_{.1}$	$n_{.2}$	…	$n_{c.}$	n

Where
n = total number of observations
I = total number of rows
c = total number of columns
n_{jk} = number of observations in category A_j and B_k (joint frequencies), j=1,..,I and k=1,..,c
$n_{j.}$ = total in row j (marginal frequencies)
$n_{.k}$ = total in column k (marginal frequencies)

CONTINGENCY TABLE: FUNDAMENTAL RELATIONSHIPS

1. **Total frequencies:**

$$\boxed{\sum_{j=1}^{l}\sum_{k=1}^{c} n_{jk} = \sum_{j=1}^{l} n_{j.} = \sum_{k=1}^{c} n_{.k} = n}$$

The sum of all joint frequencies n_{jk} equals the sum of the row totals $n_{j.}$ and the column totals $n_{.k}$, all equal to the total number of observations n.

2. **Row totals:**

$$\boxed{n_{j.} = \sum_{k=1}^{c} n_{jk} \quad \text{for each row } j}$$

Each row total $n_{j.}$ is the sum of the joint frequencies in that row.

3. **Column totals:**

$$\boxed{n_{.k} = \sum_{j=1}^{l} n_{jk} \quad \text{for each column } k}$$

Each column total $n_{.k}$ is the sum of the joint frequencies in that column.

EXAMPLE OF A CONTINGENCY TABLE

Variables:

- **Sex:** Male, Female
- **Management Level:** Junior, Mid, Senior

	Mangement Level			
Sex	**Junior**	**Mid**	**Senior**	**Total**
Male	12	20	8	40
Female	18	22	10	50
Total	30	42	18	90

Explanation:

- Rows: Sex of employees
- Columns: Management level
- Cells: Joint frequencies n_{jk}
- Row totals $n_{j.}$ and column totals $n_{.k}$ shown in the margins
- Total: n = 90

INDEPENDENCE BETWEEN TWO ATTRIBUTES

Definition:

Two attributes A and B are said to be **independent** if the joint frequencies can be expressed as the product of the corresponding marginal proportions:

$$\boxed{\frac{n_{jk}}{n} = \frac{n_{j.}}{n} \times \frac{n_{.k}}{n}}$$

Interpretation:

Independence means that the proportion of each cell is **proportional** to the product of the corresponding row and column totals.

For example, considering row j:

$$\boxed{\frac{n_{jk}}{n_{j.}} = \frac{n_{.k}}{n}}$$

Equivalently, using properties of proportions:

$$\boxed{n_{jk} = \frac{n_{j.} \times n_{.k}}{n}}$$

This expression shows that the frequency in each cell can be obtained as the product of the corresponding marginal totals divided by the grand total — the **fundamental condition of independence**.

PEARSON'S CHI-SQUARE STATISTIC

To assess independence, the **Pearson Chi-Square statistic** is computed:

$$\chi^2 = \sum_{j=1}^{l} \sum_{k=1}^{c} \frac{(n_{jk} - n^*_{jk})^2}{n^*_{jk}}$$

Interpretation:

It measures the discrepancy between the observed frequencies n_{jk} and the expected frequencies n^*_{jk} ,where

$$n^*_{jk} = \frac{n_{j.} \times n_{.k}}{n}, \quad j = 1, \ldots, l \text{ and } k = 1, \ldots, c$$

Meaning:

- The **further the Chi-Square value is from zero**, the **less credible** the assumption of independence between the two attributes becomes.

MEASURES OF ASSOCIATION

All measures listed below are based on the **Pearson Chi-Square statistic**:

$$\chi^2 = \sum_{j=1}^{l} \sum_{k=1}^{c} \frac{(n_{jk} - n_{jk}^*)^2}{n_{jk}^*}$$

Measure	Formula	Range	Interpretation
Contingency Square	$\Phi^2 = \frac{\chi^2}{n}$	$\phi^2 \geq 0$	Increases with the degree of association between variables
Contingency Coefficient	$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{\Phi^2}{\Phi^2 + 1}}$	$0 \leq C < 1$	Measures the strength of association; bounded below 1
Tschuprow's Coefficient	$T = \sqrt{\frac{\Phi^2}{\sqrt{(l-1)(c-1)}}}$	$0 \leq T \leq 1$	Adjusts for the number of rows and columns
Cramer's V Coefficient	$V = \sqrt{\frac{\Phi^2}{\min(l-1,\, c-1)}}$	$0 \leq V \leq 1$	Standardized measure of association (most used)

Interpretation:

- $\Phi^2 \rightarrow$ overall measure of association
- $C, T, V \in [0, 1]$
  - 0 → no association
  - 1 → perfect association

CALCULATING ASSOCIATION MEASURES: EXAMPLE 1

Example Table: Employees by Sex and Management Level

	Mangement Level			
Sex	**Junior**	**Mid**	**Senior**	**Total**
Male	12 (13.33)	20 (18.67)	8 (8.00)	40
Female	18 (16.67)	22 (23.33)	10 (10.00)	50
Total	30	42	18	90

Step 1: Compute expected frequencies

$$n^*_{jk} = \frac{n_{j.} \times n_{.k}}{n}$$

Example for **Male-Junior**:

$$n^*_{Male,Junior} = \frac{40 \times 30}{90} = 13.33$$

Note:
- Values in parentheses are the expected frequencies $n_{jk}*$.

CALCULATING ASSOCIATION MEASURES: EXAMPLE 1

Example Table: Employees by Sex and Management Level

	Mangement Level			
Sex	**Junior**	**Mid**	**Senior**	**Total**
Male	12 (13.33)	20 (18.67)	8 (8.00)	40
Female	18 (16.67)	22 (23.33)	10 (10.00)	50
Total	30	42	18	90

Observed n_{jk}	Expected n^*_{jk}	Contribution
12	13.33	$(12\text{-}13.33)^2/13.33 \approx 0.133$
20	18.67	$(20\text{-}18.67)^2/18.67 \approx 0.095$
8	8	$(8\text{-}8)^2/8 = 0$
18	16.67	$(18\text{-}16.67)^2/16.67 \approx 0.107$
22	23.33	$(22\text{-}23.33)^2/23.33 \approx 0.076$
10	10	$(10\text{-}10)^2/10 = 0$

Step 2: Compute Pearson Chi-Square

$$\chi^2 = \sum_{j=1}^{2}\sum_{k=1}^{3}\frac{(n_{jk}-n^*_{jk})^2}{n^*_{jk}}$$

$$\chi^2 \approx 0.133 + 0.095 + 0 + 0.107 + 0.076 + 0 = 0.411$$

CALCULATING ASSOCIATION MEASURES: EXAMPLE 1

Step 3: Compute Measures of Association

1. Contingency Square

$$\Phi^2 = \frac{\chi^2}{n} = \frac{0.411}{90} \approx 0.00457$$

2. Contingency Coefficient

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{0.411}{0.411 + 90}} \approx \sqrt{0.00457} \approx 0.0676$$

3. Tschuprow's Coefficient

$$T = \sqrt{\frac{\Phi^2}{\sqrt{(l-1)(c-1)}}} = \sqrt{\frac{0.00457}{\sqrt{(2-1)(3-1)}}} = \sqrt{\frac{0.00457}{\sqrt{2}}} \approx 0.048$$

4. Cramer's V Coefficient

$$V = \sqrt{\frac{\Phi^2}{\min(l-1,\, c-1)}} = \sqrt{\frac{0.00457}{\min(1,2)}} = \sqrt{0.00457} \approx 0.0676$$

Interpretation:
All measures are **very close to zero**, indicating a **very weak association** between Sex and Management Level in this example.

ROUGH GUIDELINES FOR STRENGTH OF ASSOCIATION (CRAMER'S V / TSCHUPROW)

Value	Strength
0.00 – 0.10	Very weak
0.10 – 0.30	Weak
0.30 – 0.50	Moderate
0.50 – 0.70	Strong
0.70 – 1.00	Very strong

Note:

- The table above presents commonly used guidelines for interpreting the strength of association for **Cramer's V** and **Tschuprow's T.**
- For the **Contingency Coefficient C**, the maximum value depends on the table size, so interpretation should be done carefully.
- These ranges are **guidelines**, not strict rules; some fields may adopt slightly different cutoffs.

2×2 CONTINGENCY TABLES: DEFINITION AND EXAMPLE

Definition

- A 2×2 contingency table summarizes the frequency of observations for **two categorical variables**, each with **two categories**.
- It helps analyze the **association between variables**.

General Example of a 2×2 Contingency Table

- Consider two categorical variables, A and B, each with two categories.

	B		
A	**+**	**-**	**Total**
+	a	b	a+b
-	c	d	c+d
Total	a+c	b+d	n

Where
n = total number of observations
l = total number of rows = 2
c = total number of columns =2
The symbols + and – denote the presence and absence of the attribute, respectively.

EXAMPLE OF A 2×2 CONTINGENCY TABLE

Consider two categorical variables:

- **A:** Smoking status (+ = Smoker, – = Non-Smoker)
- **B:** Lung disease (+ = Disease, – = No Disease)

	Disease		
Smoker	**Yes**	**No**	**Total**
Yes	30	20	50
No	10	40	50
Total	40	60	100

SIMPLIFIED PEARSON'S CHI-SQUARE STATISTIC FOR 2×2 TABLES

For 2×2 contingency tables, **Pearson's chi-square statistic** has a simplified formula:

$$\chi^2 = \frac{n\,(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Where

- n = a+b+c+d = total sample size
- a, b, c ,d = observed frequencies

Note:

- This formula **avoids computing expected frequencies explicitly** and is valid **only for 2×2 tables**.

ASSOCIATION MEASURES FOR 2×2 CONTINGENCY TABLES

Measure	Formula	Range	Interpretation
Phi Coefficient (Φ)	$\boxed{\Phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}}$	$-1 \leq \Phi \leq 1$	Measures strength and direction of association; 0 = no association, positive/negative indicates direction
Yule's Q	$\boxed{Q = \frac{ad - bc}{ad + bc}}$	$-1 \leq Q \leq 1$	Measures association for binary variables; 0 = no association, sign indicates direction; symmetric measure
Yule's Y (Colligation Coefficient)	$\boxed{Y = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}}$	$-1 \leq Y \leq 1$	Alternative measure of association for 2×2 tables; reduces sensitivity to extreme values; 0 = no association, sign indicates direction

YULE'S COEFFICIENTS (2×2 TABLES)

Relationship between Q and Y:

$$\boxed{Q = \frac{2Y}{1+Y^2}}$$

Notes:

- **Yule's Q and Yule's Y are related**: both indicate the same direction of association.
- **Yule's Y reduces sensitivity to extreme frequencies**, i.e., when some cells have very small or very large counts.
- **Sign (+/–) indicates direction**, absolute value indicates strength.

DIFFERENCES BETWEEN PHI, YULE'S Q AND YULE'S Y

Coefficient	Direction	Strength	Notes / Purpose
Phi (Φ)	Yes, sign indicates direction (+/-)	Yes, magnitude indicates strength	Measures linear association for **binary variables**, similar to correlation; sensitive to marginal distributions
Yule's Q	Yes, sign indicates direction (+/-)	Yes, magnitude indicates strength	Measures **association (dependence)** between two binary variables; symmetric; less sensitive to marginal totals than Φ in some cases
Yule's Y (Colligation Coefficient)	Yes, sign indicates direction (+/-)	Yes, magnitude indicates strength	Alternative to Q; reduces influence of extreme frequencies; also measures association (strength & direction)

INTERPRETATION GUIDELINES FOR 2×2 ASSOCIATION MEASURES (Φ, YULE'S Q, YULE'S Y)

Measure	Value (absolute)	Strength of Association (approx.)
Phi (Φ)	0 – 0.1	Negligible / very weak
	0.1 – 0.3	Weak
	0.3 – 0.5	Moderate
	0.5 – 0.7	Strong
	0.7 – 1	Very strong
Yule's Q or Y	0 – 0.3	Weak
	0.3 – 0.5	Moderate
	0.5 – 0.7	Strong
	0.7 – 1	Very strong

Note:

- The table refers to the absolute values of the association measures. The sign of the coefficient (+ or –) indicates the direction of the association: positive values represent a positive association, negative values represent a negative association.

CONTINGENCY SQUARE (Φ^2) VS. PHI (Φ) FOR 2×2 TABLES

1. **Phi Coefficient (Φ)**

$$\boxed{\Phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}}$$

- Expressed directly in terms of **cell frequencies**.
- Values: $-1 \le \Phi \le 1$
- Measures **both strength and direction** of association.

2. **Contingency Square (Φ^2)**

$$\boxed{\Phi^2 = \frac{\chi^2}{n}}$$

- Derived from **Pearson's chi-square statistic**.
- Values: $0 \le \Phi^2 \le 1$ (for 2×2 tables, maximum can reach 1)
- Measures **strength only**; loses direction information.

For 2×2 tables:

$$\Phi^2 = (\Phi)^2$$

That is, the **square of the Phi coefficient** equals the contingency square.
So Φ^2 gives the **magnitude**, while Φ also gives the **direction** of association.

CALCULATING ASSOCIATION MEASURES: EXAMPLE 2

Example Table: Smoker and Disease

a = 30, b = 20, c = 10, d = 40

	Disease		
Smoker	**Yes**	**No**	**Total**
Yes	30	20	50
No	10	40	50
Total	40	60	100

Step 1: Pearson's Chi-Square

$$\chi^2 = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Substituindo:

$$\chi^2 = \frac{100(30 \cdot 40 - 20 \cdot 10)^2}{(30+20)(10+40)(30+10)(20+40)}$$

$$\chi^2 = \frac{100(1200 - 200)^2}{50 \cdot 50 \cdot 40 \cdot 60} = \frac{100(1000)^2}{6,000,000} = \frac{100,000,000}{6,000,000} \approx 16.67$$

CALCULATING ASSOCIATION MEASURES: EXAMPLE 2

Example Table: Smoker and Disease

a = 30, b = 20, c = 10, d = 40

	Disease		
Smoker	**Yes**	**No**	**Total**
Yes	30	20	50
No	10	40	50
Total	40	60	100

Step 2: Phi Coefficient (Φ)

$$\Phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

$$\Phi = \frac{30 \cdot 40 - 20 \cdot 10}{\sqrt{50 \cdot 50 \cdot 40 \cdot 60}} = \frac{1200 - 200}{\sqrt{6,000,000}} = \frac{1000}{2449.49} \approx 0.408$$

CALCULATING ASSOCIATION MEASURES: EXAMPLE 2

Example Table: Smoker and Disease

a = 30, b = 20, c = 10, d = 40

	Disease		
Smoker	**Yes**	**No**	**Total**
Yes	30	20	50
No	10	40	50
Total	40	60	100

Step 3: Yule's Q

$$Q = \frac{ad - bc}{ad + bc} = \frac{1200 - 200}{1200 + 200} = \frac{1000}{1400} \approx 0.714$$

Step 4: Yule's Y

$$Y = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}} = \frac{\sqrt{1200} - \sqrt{200}}{\sqrt{1200} + \sqrt{200}} \approx \frac{34.64 - 14.14}{34.64 + 14.14} \approx \frac{20.50}{48.78} \approx 0.420$$

Interpretation:

- **χ^2 ≈ 16.67** → significant association
- **Φ ≈ 0.408** → moderate positive association
- **Yule's Q ≈ 0.714** → strong positive association (focuses on odds ratio)
- **Yule's Y ≈ 0.420** → moderate positive association, less extreme than Q

All measures indicate that smoking is positively associated with lung disease.

COMPARISON OF ASSOCIATION MEASURES: GENERAL VS 2×2 TABLES

Notes:

- **General tables (l×c):** Only measure strength; no direction.
- **2×2 tables:** Measures give both strength and direction because variables are binary.

Measure	Used For	Range	Direction	Interpretation
Φ^2 (Contingency Square)	l×c tables	$\Phi^2 \geq 0$	No	Strength of association only
Contingency Coefficient (C)	l×c tables	$0 \leq C < 1$	No	Standardized strength; depends on table size
Tschuprow's T	l×c tables	$0 \leq T \leq 1$	No	Adjusts for table dimensions; strength only
Cramer's V	l×c tables	$0 \leq V \leq 1$	No	Standardized; widely used; strength only
Phi (Φ)	2×2 tables	$-1 \leq \Phi \leq 1$	Yes	Strength and direction; correlation-like
Yule's Q	2×2 tables	$-1 \leq Q \leq 1$	Yes	Strength and direction; symmetric association
Yule's Y	2×2 tables	$-1 \leq Y \leq 1$	Yes	Strength and direction; less sensitive to extreme frequencies

THANKS!

Questions?