STATISTICAL LABORATORY

Lecture 7: Contingency Tables and Measures Association (Theory and Practice)

Yiu Lim LUI, Allen

Lisbon School of Economics & Management
Universidade de Lisboa

allenlui@iseg.ulisboa.pt

Applied Mathematics for Economics and Management
1st Year / 1st Semester – 2026/2027

Reminder

- The group report must be submitted in PDF format to the lecture instructor via email at allenlui@iseg.ulisboa.pt by 20:00 on November 20, 2026.
- For detailed group project guidelines, please consult the Fenix portal.

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

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 $\mathbf{A_1}, \mathbf{A_2}, \ldots, \mathbf{A_I}$ and variable **B** have categories $\mathbf{B_1}, \mathbf{B_2}, \ldots, \mathbf{B_c}$.

	$\mathbf{B_1}$	$\mathbf{B_2}$	$---$	$\mathbf{B_c}$	**Totals**
$\mathbf{A_1}$	n_{11}	n_{12}	$\ldots$	n_{1c}	$n_{1.}$
$\mathbf{A_2}$	n_{21}	n_{22}	$\ldots$	n_{2c}	$n_{2.}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
$\mathbf{A_I}$	n_{I1}	n_{I2}	$\ldots$	n_{Ic}	$n_{I.}$
Totals	$n_{.1}$	$n_{.2}$	$\ldots$	$n_{.c}$	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), $j = 1, \ldots, I$ and $k = 1, \ldots, c$

$n_{j.}$ = total in row j (marginal frequencies)

$n_{.k}$ = total in column k (marginal frequencies)

Notation

- n_{jk}: The observed frequency where j is the row index and k is the column index.
- In math notation, the dot replaces the index you add up.
  - $n_{j.}$: The sum of the j-th row (the dot indicates summation over all columns k).
  - $n_{.k}$: The sum of the k-th column (the dot indicates summation over all rows j).

Simple example

Suppose we survey 100 random trips, comparing weather condition (Rows: A_1 = Sunny, A_2 = Rainy) and primary transportation method (Columns: B_1 = Walking, B_2 = Metro, B_3 = Bus).

	Walking (B_1)	**Metro (B_2)**	**Bus (B_3)**	**Totals**
Sunny (A_1)	30	20	10	60
Rainy (A_2)	5	25	10	40
Totals	35	45	20	100

- Key takeaways
  - $n = 100$ total observations
  - $n_{11} = 30$ (sunny & walking)
  - $n_{22} = 25$ (rainy & metro)
  - Row totals give weather frequencies; column totals give transit preferences.

Interpreting the contingency table

- **What the table shows:** Cross-classifies 100 recorded trips by two categorical variables: weather condition ($A_1 =$ Sunny, $A_2 =$ Rainy) and preferred transportation ($B_1 =$ Walking, $B_2 =$ Metro, $B_3 =$ Bus).
- **Joint frequencies:** Cell numbers show simultaneous occurrences.
  - 30 trips (n_{11}) occurred during sunny weather and walking.
  - 25 trips (n_{22}) occurred during rainy weather and metro use.
- **Marginal frequencies (totals):**
  - Row totals show weather frequencies (60 sunny, 40 rainy).
  - Column totals show transport frequencies (35 walking, 45 metro, 20 bus).
- **Grand total ($n = 100$):** Located at the bottom-right corner, representing the total sample size.

Contingency table: fundamental relationships

1. **Total frequencies:**

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

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:**

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

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

3. **Column totals:**

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

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

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:

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

- **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:

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

- **Equivalently, using properties of proportions:**

$$n_{jk} = \frac{n_{j.} \times n_{.k}}{n} \tag{1}$$

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**.

Previous weather example

- If weather and transportation choices were truly independent, it would mean that people's travel habits do not change based on the weather.
- Consequently, regardless of the weather, we should expect the same relative frequency (proportion) of walking, taking the metro, or taking the bus.
- This concept is captured directly by the formula on the previous slide:

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

- Left-Hand Side ($\frac{n_{jk}}{n_{j.}}$): This is the conditional proportion–the fraction of people choosing a specific transport method (k) within a specific weather condition (j).
- Right-Hand Side ($\frac{n_{.k}}{n}$): This is the overall marginal proportion–the baseline fraction of people choosing that transport method across the entire city, ignoring the weather entirely.

Previous weather example (Continued)

- For example, looking at walking ($k = 1$) on sunny days ($j = 1$):

$$\frac{n_{11}}{n_{1.}} = \frac{30}{60} = 0.50 \quad (50\%)$$

- For walking overall ($k = 1$):

$$\frac{n_{.1}}{n} = \frac{35}{100} = 0.35 \quad (35\%)$$

Previous weather example (Continued)

- Under independence, these two values must be equal. However, in our real-world data:

$$0.50 \neq 0.35$$

- Because 50% of people walk when it is sunny, compared to only 35% overall, the proportions shift.
- This confirms visually and mathematically that weather alters travel behavior, proving that weather and transportation are not independent.

Pearson's chi-square statistic

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

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

(Chi-Square is pronounced as "Kai-square" in English.)

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

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

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

Pearson's chi-square statistic (Continued)

- n_{jk} (Observed): The actual number of times a specific combination happened in your data (e.g., how many people walked on a sunny day).
- n^*_{jk} (Expected): The number of observations you would expect in that cell if the two variables were completely independent (calculated via $\frac{n_{j.} \times n_{.k}}{n}$ from Equation (1)).
- If two variables were completely independent, we should expect $n_{jk} - n^*_{jk} = 0$.
- Dividing by n^*_{jk} standardizes the deviation, turning an absolute count difference into a relative, proportional measure.
- Squaring $((n_{jk} - n^*_{jk})^2)$ ensures no cancellation.

Measures of association

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

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

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

- I: number of rows in the contingency table
- c: number of columns in the contingency table

Interpretation summary:
- Φ^2 is an overall measure of association
- $C, T, V \in [0, 1]$
- $0 \rightarrow$ no association
- $1 \rightarrow$ perfect association

Measures of association (Continued)

Why use adjusted association measures?

1. The Sample Size Problem $\rightarrow$ Contingency Square (Φ^2)
   - The value of χ^2 scales directly with the sample size n. If you double your sample size but the relationship remains exactly the same, your χ^2 value doubles.
   - Dividing by n ($\Phi^2 = \frac{\chi^2}{n}$) removes the effect of the sample size. However, it has a major drawback: it has no upper bound ($\Phi^2 \geq 0$). For tables larger than 2×2, Φ^2 can easily exceed 1, making it hard to interpret how "strong" the relationship is.

Measures of association (Continued)

2 The Upper Bound Problem $\rightarrow$ Contingency Coefficient (C)

- Karl Pearson created C to bound the value between 0 and 1, making it look like a standard correlation coefficient.
- The New Issue: C can never actually reach 1, even if there is a perfect relationship. The maximum possible value it can reach depends entirely on the size of the table (e.g., in a 2×2 table, the maximum C is roughly 0.707). You cannot use it to compare tables of different sizes.

Measures of association (Continued)

3 The Table Dimension Problem $\rightarrow$ Tschuprow's T vs. Cramer's V

To fix the issues of both Φ^2 and C, statisticians looked for ways to force the value to be exactly between 0 and 1, regardless of how many rows (I) and columns (c) the table has.

- **Tschuprow's** T: Divides by the geometric mean of the dimensions ($\sqrt{(I-1)(c-1)}$). It hits 1 only if the table is perfectly square ($I = c$) and the association is perfect. If the table is rectangular (e.g., 2×3), T can never reach 1.
- **Cramer's** V: Divides by the absolute minimum possible dimension ($\min(I-1, c-1)$). Cramer's V is guaranteed to reach exactly 1 for perfect association in any table shape (square or rectangular).

Example : Symmetric Perfect Association (2x2)

Consider a strict diagnostic health check with zero false positives or false negatives ($n = 100$):

	Sick	**Healthy**	**Row Total**
Positive Test	50	0	50
Negative Test	0	50	50
Column Total	50	50	$n = 100$

- Under independence, the expected frequency for each cell is $n^*_{jk} = 25$.
- This yields a raw Pearson statistic of:

$$\chi^2 = \sum \frac{(n_{jk} - n^*_{jk})^2}{n^*_{jk}} = 4 \times \frac{(\pm 25)^2}{25} = 100$$

- The contingency square is: $\Phi^2 = \frac{\chi^2}{n} = \frac{100}{100} = 1$

Example: Calculating C, T, and V

Using $\chi^2 = 100$, $n = 100$, and table dimensions $I = 2, c = 2$:

1. **Contingency Coefficient (C):** Flawed upper bound.

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{100}{100 + 100}} = \sqrt{0.5} \approx 0.707$$

2. **Tschuprow's T:** Hits 1.0 because the table is perfectly square ($I = c$).

$$T = \sqrt{\frac{\Phi^2}{\sqrt{(I-1)(c-1)}}} = \sqrt{\frac{1}{\sqrt{(2-1)(2-1)}}} = 1.0$$

3. **Cramer's V:** Standardized approach using the minimum dimension.

$$V = \sqrt{\frac{\Phi^2}{\min(I-1, c-1)}} = \sqrt{\frac{1}{\min(1,1)}} = 1.0$$

Association calculating measures: Example I

Example Table: Employees by Sex and Management Level

Sex \ Level	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

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

- **Note:** Values in parentheses are expected frequencies n^*_{jk}.
- For Male-Junior: $n^*_{\text{Male,Junior}} = \frac{40 \times 30}{90} = 13.33$.

Association calculating measures: Example I (Cont.)

Step 2: Compute Pearson Chi-Square

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

$$\chi^2 = \sum_{j=1}^{2}\sum_{k=1}^{3}\frac{(n_{jk}-n^*_{jk})^2}{n^*_{jk}} \approx 0.133+0.095+0+0.107+0.076+0 = 0.411$$

Association calculating measures: Example I (Conclusion)

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:**

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

3. **Tschuprow's Coefficient:**

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

4. **Cramer's V Coefficient:**

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

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

Notes:

- The table 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 table size, so interpret with care.
- These ranges are guidelines, not strict rules; cutoffs can vary by field.

Exercise 4.4

The table below presents the election results in a given district:

Political Party	**A**	**B**	**C**
Employed	393	232	152
Unemployed	55	28	20

- Using two appropriate measures, assess the existence of an association between employment status and voting choice. Justify your answer.

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 Structure:**

A \ B	$+$	$-$	**Totals**
$+$	a	b	$a+b$
$-$	c	d	$c+d$
Totals	$a+c$	$b+d$	n

Where: $n =$ total number of observations, $I = 2$ rows, $c = 2$ columns. The symbols $+$ and $-$ denote the presence and absence of the attribute.

Example of a 2×2 contingency table

Consider two categorical variables:

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

Smoking \ Disease	**Yes (+)**	**No (−)**	**Totals**
Yes (+)	30	20	50
No (−)	10	40	50
Totals	60	40	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)	$\frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$	$-1 \leq \Phi \leq 1$	Measures strength and direction; 0 = no assoc., positive/negative indicates direction
Yule's Q	$\frac{ad - bc}{ad + bc}$	$-1 \leq Q \leq 1$	Binary association; 0 = no assoc., sign indicates direction; symmetric
Yule's Y	$\frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$	$-1 \leq Y \leq 1$	Reduces sensitivity to extreme values; 0 = no assoc., sign indicates direction

Key takeaways:
- All three measures indicate both the **strength** and **direction** ($+$ or $-$) of the association.

Yule's coefficients (2×2 tables)

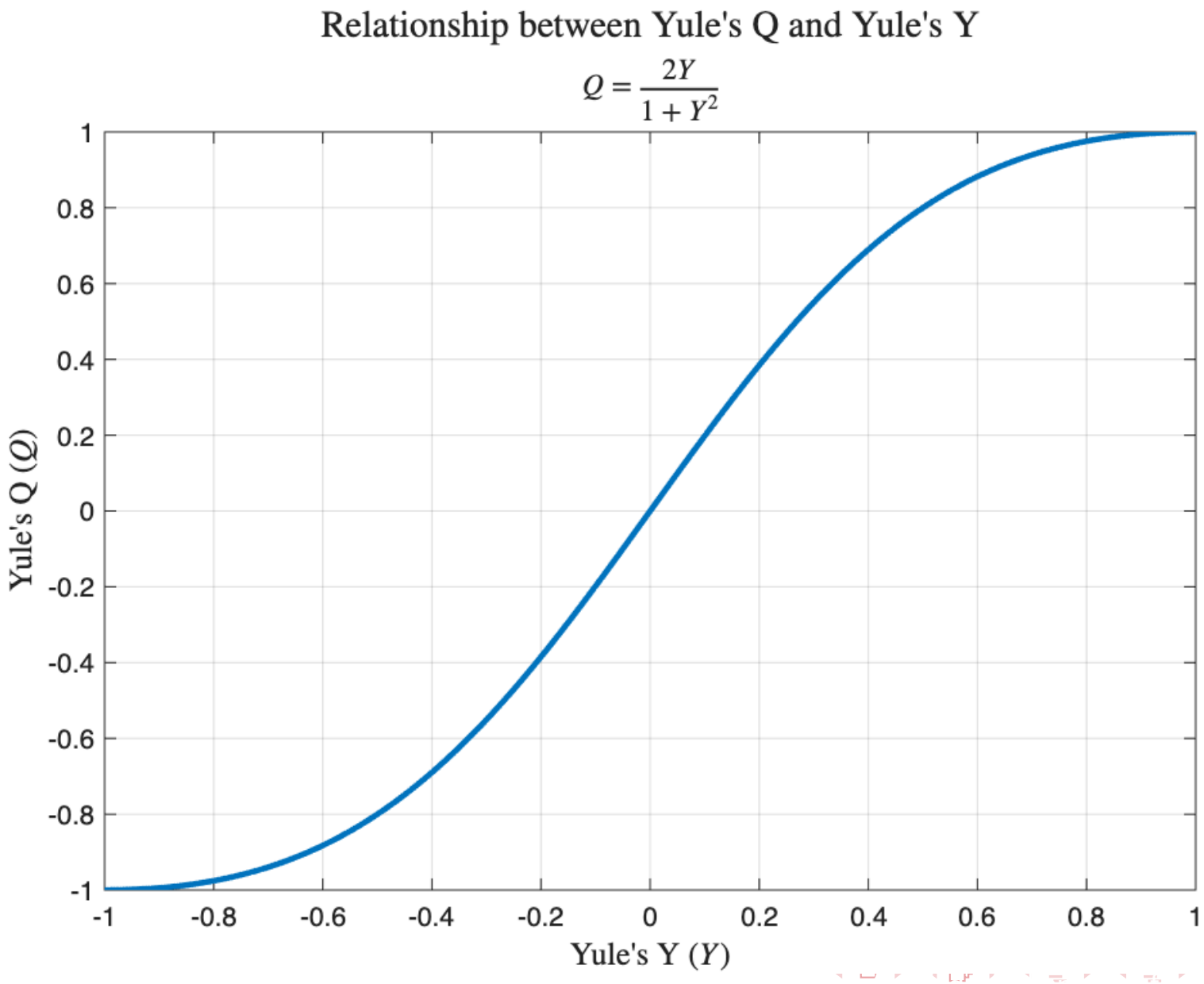
Relationship between Q and Y:

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

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.

Relationship between Q and Y



Differences between Phi, Yule's Q and Yule's Y

Coefficient	Direction	Strength	Notes / Purpose
Phi (Φ)	Yes ($+/-$)	Yes	Measures linear association for binary variables, similar to correlation; sensitive to marginal distributions
Yule's Q	Yes ($+/-$)	Yes	Measures association between two binary variables; symmetric; less sensitive to marginal totals
Yule's Y	Yes ($+/-$)	Yes	Alternative to Q; reduces influence of extreme frequencies; measures association strength & direction

Interpretation guidelines for 2×2 association measures

Measure	Value (absolute)	Strength of Association (approx.)
Phi (Φ)	$0.0 - 0.1$	Negligible / very weak
	$0.1 - 0.3$	Weak
	$0.3 - 0.5$	Moderate
	$0.5 - 0.7$	Strong
	$0.7 - 1.0$	Very strong
Yule's Q or Y	$0.0 - 0.3$	Weak
	$0.3 - 0.5$	Moderate
	$0.5 - 0.7$	Strong
	$0.7 - 1.0$	Very strong

Note: Ranges refer to absolute values. The sign ($+$ or $-$) indicates the direction of the association (positive vs. negative).

Contingency square (Φ^2) vs. Phi (Φ) for 2×2 tables

1. **Phi Coefficient (Φ):**

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

Range: $-1 \leq \phi \leq 1$. Measures both strength and direction.

2. **Contingency Square (Φ^2):**

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

Range: $0 \leq \phi^2 \leq 1$. Measures strength only; loses direction.

Key Relationship for 2×2 Tables:
- $\Phi^2 = (\Phi)^2 \quad \rightarrow \quad \Phi^2$ gives the magnitude, while Φ gives the direction.

Association calculating measures: Example 2

Example Table: Smoker and Disease
$(a = 30, b = 20, c = 10, d = 40, n = 100)$

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

Step 1: Pearson's Chi-Square

$$\chi^2 = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{100(30 \cdot 40 - 20 \cdot 10)^2}{50 \cdot 50 \cdot 40 \cdot 60} \approx 16.67$$

Association calculating measures: Example 2 (Cont.)

Example Table: Smoker and Disease

$(a = 30, b = 20, c = 10, d = 40, n = 100)$

Step 2: Phi Coefficient (Φ)

$$\Phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} = \frac{30 \cdot 40 - 20 \cdot 10}{\sqrt{50 \cdot 50 \cdot 40 \cdot 60}} = \frac{1000}{\sqrt{6,000,000}} \approx 0.408$$

Step 3: Yule's Q & Yule's Y

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

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

Association calculating measures: Example 2 (Conclusion)

Interpretation Summary for Example 2:

- $\chi^2 \approx 16.67 \rightarrow$ significant association
- $\Phi \approx 0.408 \rightarrow$ moderate positive association
- Yule's $Q \approx 0.714 \rightarrow$ strong positive association (focuses on odds ratio)
- Yule's $Y \approx 0.420 \rightarrow$ moderate positive association, less extreme than Q

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

Sensitivity to Marginal distributions

Imagine a clinical trial testing a treatment. In both scenarios below, the treatment's true effectiveness is identical: 80% of treated patients recover, while only 20% of untreated patients recover.

Scenario 1: Perfectly Balanced Groups ($n = 100$)

Group	Recovered	Not Recovered	Total
Treated	**40** (80%)	10	**50**
Untreated	10	**40** (80%)	**50**
Total	**50**	**50**	**100**

- **Phi Coefficient (Φ):** High sensitivity to equal distributions.

$$\Phi = \frac{(40 \times 40) - (10 \times 10)}{\sqrt{50 \times 50 \times 50 \times 50}} = \frac{1500}{2500} = \mathbf{0.60}$$

- **Yule's Q:**

$$Q = \frac{(40 \times 40) - (10 \times 10)}{(40 \times 40) + (10 \times 10)} = \frac{1500}{1700} \approx \mathbf{0.88}$$

Sensitivity to Marginal distributions (Continued)

Scenario 2: Severe Sample Size Imbalance ($n = 100$)

90 Treated patients vs. only 10 Untreated patients. True clinical risk unchanged.

Group	Recovered	Not Recovered	Total
Treated	**72** (80%)	18	**90**
Untreated	2	**8** (80%)	**10**
Total	**74**	**26**	**100**

- **Phi Coefficient (Φ):** Drops heavily due to skewed margins.

$$\Phi = \frac{(72 \times 8) - (18 \times 2)}{\sqrt{90 \times 10 \times 74 \times 26}} = \frac{540}{\sqrt{1,731,600}} \approx \mathbf{0.41} \quad (\downarrow 32\%)$$

- **Yule's Q:** Remains completely constant and stable.

$$Q = \frac{(72 \times 8) - (18 \times 2)}{(72 \times 8) + (18 \times 2)} = \frac{540}{612} \approx \mathbf{0.88}$$

Comparison of association measures: general vs 2×2 tables

Measure	Used For	Range	Direction	Interpretation
Φ^2 (Cont. Square)	$I \times c$ tables	$\Phi^2 \geq 0$	No	Strength of association only
Cont. Coefficient (C)	$I \times c$ tables	$0 \leq C < 1$	No	Standardized strength; depends on table size
Tschuprow's T	$I \times c$ tables	$0 \leq T \leq 1$	No	Adjusts for table dimensions; strength only
Cramer's V	$I \times c$ 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 extremes

Exercise 4.3

Consider the following table on employment status by gender in a given region:

	Employed	**Unemployed**
Women	100	55
Men	120	35

a) Assess the existence of an association between gender and employment status using an appropriate measure.

b) Identify, in statistical terms, the frequencies corresponding to the following expressions:
b1) 100 / 155 **b2)** 35 / 310 **b3)** 155 / 310

Exercise 4.4 (Suggested Answers)

Employment Status	**Party A**	**Party B**	**Party C**	**Row Totals**
Employed	393	232	152	**777**
Unemployed	55	28	20	**103**
Column Totals	**448**	**260**	**172**	$n = 880$

Exercise 4.4 (Suggested Answers)

Under the fundamental condition of independence, the expected frequency for each cell is given by:

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

Computing each cell expectation n^*_{jk}:

- **Employed & Party A:** $n^*_{11} = \frac{777 \times 448}{880} \approx 395.564$
- **Employed & Party B:** $n^*_{12} = \frac{777 \times 260}{880} \approx 229.568$
- **Employed & Party C:** $n^*_{13} = \frac{777 \times 172}{880} \approx 151.868$
- **Unemployed & Party A:** $n^*_{21} = \frac{103 \times 448}{880} \approx 52.436$
- **Unemployed & Party B:** $n^*_{22} = \frac{103 \times 260}{880} \approx 30.432$
- **Unemployed & Party C:** $n^*_{23} = \frac{103 \times 172}{880} \approx 20.132$

Exercise 4.4 (Suggested Answers)

Using Pearson's chi-square formula to measure the discrepancy between observed (n_{jk}) and expected (n^*_{jk}) frequencies:

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

$$\chi^2 = \frac{(393 - 395.564)^2}{395.564} + \frac{(232 - 229.568)^2}{229.568} + \frac{(152 - 151.868)^2}{151.868} + \frac{(55 - 52.436)^2}{52.436} + \frac{(28 - 30.432)^2}{30.432} + \frac{(20 - 20.132)^2}{20.132}$$

$$= 0.0166 + 0.0258 + 0.0001 + 0.1254 + 0.1944 + 0.0009 \approx \mathbf{0.3630}$$

Exercise 4.4 (Suggested Answers)

$$\Phi^2 = \frac{\chi^2}{n} = \frac{0.3630}{880} \approx \mathbf{0.0004125}$$

1. **Tschuprow's Coefficient (T):**

$$T = \sqrt{\frac{\Phi^2}{\sqrt{(I-1)(c-1)}}} = \sqrt{\frac{0.0004125}{\sqrt{(2-1)(3-1)}}} = \sqrt{\frac{0.0004125}{\sqrt{2}}} \approx \mathbf{0.0171}$$

2. **Cramer's V Coefficient (V):**

$$V = \sqrt{\frac{\Phi^2}{\min(I-1, c-1)}} = \sqrt{\frac{0.0004125}{\min(1,2)}} = \sqrt{\frac{0.0004125}{1}} \approx \mathbf{0.0203}$$

According to standard interpretation guidelines ($0.00 - 0.10$ indicates a *very weak* association), both coefficients confirm that there is a **very weak association** between employment status and voting choice.

Exercise 4.3 (Suggested Answers)

a) Using the formula for Yule's Q:

$$Q = \frac{ad - bc}{ad + bc}$$

Substituting the cell frequencies ($a = 100$, $b = 55$, $c = 120$, $d = 35$):

$$ad = 100 \times 35 = 3500$$

$$bc = 55 \times 120 = 6600$$

$$Q = \frac{3500 - 6600}{3500 + 6600} = \frac{-3100}{10100} \approx \mathbf{-0.307}$$

Exercise 4.3 (Suggested Answers)

$$Y = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$$

Substituting $\sqrt{ad} = \sqrt{3500} \approx 59.161$ and $\sqrt{bc} = \sqrt{6600} \approx 81.240$:

$$\begin{aligned} Y &= \frac{59.161 - 81.240}{59.161 + 81.240} \\ &= \frac{-22.079}{140.401} \approx \mathbf{-0.157} \end{aligned}$$

Exercise 4.3 (Suggested Answers)

b) b1) Conditional Relative Frequency of being employed given that the person is a woman.
b2) Joint Relative Frequency of being a man and unemployed.
b3) Marginal Relative Frequency of women.

THANKS!

Questions?