UNIVERSITÁ DEGLI STUDI DI TERAMO

# Analisi dell'associazione tra due caratteri: indipendenza e dipendenza

# Introduzione

- Analisi univariata, bivariata, multivariata
- Analizzare le relazioni tra i caratteri, per cercare di "prevedere" il valore (sconosciuto) di una variabile a partire da quello (conosciuto) di un'altra
- Distribuzioni doppie di frequenze (tabelle doppie)
- Associazione tra caratteri (dipendenza, indipendenza, ecc..)
- Il $\chi^2$ (Chi-quadrato)

# Distribuzione doppia - 1

## a) Distribuzione doppia di frequenze

| | Maschi | Femmine | Totale |
|---|---|---|---|
| Economia | 75 | 45 | 120 |
| Statistica | 35 | 65 | 100 |
| Matematica | 60 | 20 | 80 |
| Totale | 170 | 130 | 300 |

$n_{i.}$

Totale di riga

$n_{.j}$

Totale di colonna

Frequenza assoluta congiunta

$n_{ij}$

$n_{..}$

# Distribuzione doppia - 2

**a) Distribuzione doppia di frequenze**

| | Maschi | Femmine | Totale |
|---|---|---|---|
| Economia | 75 | 45 | 120 |
| Statistica | 35 | 65 | 100 |
| Matematica | 60 | 20 | 80 |
| Totale | 170 | 130 | 300 |

**Distribuzione di Y condizionata a $X_1$**

**Distribuzione Marginale di Y**

**Distribuzione di X condizionata a $Y_1$**

**Distribuzione Marginale di X**

# INDIPENDENZA di X da Y

## a) Frequenze assolute

| | Carattere Y | | | |
|---|---|---|---|---|
| Carattere X | | $Y_1$ | $Y_2$ | Totale |
| | $X_1$ | 10 | 20 | 30 |
| | $X_2$ | 20 | 40 | 60 |
| | $X_3$ | 30 | 60 | 90 |
| | Totale | 60 | 120 | 180 |

## b) Frequenze relative del carattere X condizionato ad Y

| | Carattere Y | | | |
|---|---|---|---|---|
| Carattere X | | $Y_1$ | $Y_2$ | Totale |
| | $X_1$ | 0,167 | 0,167 | 0,167 |
| | $X_2$ | 0,333 | 0,333 | 0,333 |
| | $X_3$ | 0,500 | 0,500 | 0,500 |
| | Totale | 1,000 | 1,000 | 1,000 |

**Il carattere X si dirà indipendente dal carattere Y se tutte le distribuzioni relative condizionate risultano uguali tra loro e uguali alla distribuzione marginale (e dunque, al variare della modalità Y la distribuzione relativa di X è la medesima).**

# INDIPENDENZA di Y da X

### a) Frequenze assolute

| | | Carattere Y | | |
|---|---|---|---|---|
| | | $Y_1$ | $Y_2$ | Totale |
| Carattere X | $X_1$ | 10 | 20 | 30 |
| | $X_2$ | 20 | 40 | 60 |
| | $X_3$ | 30 | 60 | 90 |
| | Totale | 60 | 120 | 180 |

### b) Frequenze relative del carattere Y condizionato ad X

| | | Carattere Y | | |
|---|---|---|---|---|
| | | $Y_1$ | $Y_2$ | Totale |
| Carattere X | $X_1$ | 0,333 | 0,667 | 1,000 |
| | $X_2$ | 0,333 | 0,667 | 1,000 |
| | $X_3$ | 0,333 | 0,667 | 1,000 |
| | Totale | 0,333 | 0,667 | 1,000 |

**Il carattere Y si dirà indipendente dal carattere X se tutte le distribuzioni relative condizionate risultano uguali tra loro e uguali alla distribuzione marginale (e dunque, al variare della modalità X la distribuzione relativa di Y è la medesima).**

# INDIPENDENZA

E' possibile dimostrare che se il carattere X è indipendente dal carattere Y, allora vale anche la relazione contraria: anche il carattere Y sarà indipendente dal carattere X.

Pertanto: due caratteri X ed Y si diranno indipendenti se le distribuzioni relative condizionate di un carattere rispetto alle modalità dell'altro sono uguali.

$$n'_{ij} = \frac{n_{i.} \cdot n_{.j}}{n}$$

**Frequenze teoriche di indipendenza**

# INDIPENDENZA

### a) Frequenze assolute

| | | Carattere Y | | |
|---|---|---|---|---|
| | | $Y_1$ | $Y_2$ | Totale |
| Carattere X | $X_1$ | 10 | 20 | 30 |
| | $X_2$ | 20 | 40 | 60 |
| | $X_3$ | 30 | 60 | 90 |
| | Totale | 60 | 120 | 180 |

$$n'_{11} = \frac{n_{1.} \cdot n_{.1}}{n} = \frac{30 \cdot 60}{180} = \frac{1.800}{180} = 10$$

$$n'_{32} = \frac{n_{3.} \cdot n_{.2}}{n} = \frac{90 \cdot 120}{180} = \frac{10.800}{180} = 60$$

**Ogni volta che non troviamo questa situazione**

↓

**Dipendenza**

# Dipendenza perfetta di Y da X

**Un carattere Y dipende perfettamente da X se ad ogni modalità di X è associata una ed una sola modalità del carattere Y**

| | | Carattere Y | | |
|---|---|---|---|---|
| | | $Y_1$ | $Y_2$ | Totale |
| Carattere X | $X_1$ | 0 | 20 | 20 |
| | $X_2$ | 20 | 0 | 20 |
| | $X_3$ | 0 | 60 | 60 |
| | Totale | 20 | 80 | 100 |

**Se $X = X_1$ → $\boxed{Y = Y_2}$**

**Se $X = X_2$ → $\boxed{Y = Y_1}$**

**Se $X = X_3$ → $\boxed{Y = Y_2}$**

**La relazione di dipendenza non è biunivoca!!!**

# Dipendenza perfetta di X da Y

**Un carattere X dipende perfettamente da Y se ad ogni modalità di Y è associata una ed una sola modalità del carattere X**

| | Carattere Y | | | | | |
|---|---|---|---|---|---|---|
| Carattere X | | $Y_1$ | $Y_2$ | $Y_3$ | $Y_4$ | Totale |
| | $X_1$ | 20 | 0 | 0 | 0 | 20 |
| | $X_2$ | 0 | 20 | 0 | 0 | 20 |
| | $X_3$ | 0 | 0 | 30 | 60 | 90 |
| | Totale | 20 | 20 | 30 | 60 | 130 |

**Se $Y = Y_1$ ⟶ $\boxed{X = X_1}$**

**Se $Y = Y_2$ ⟶ $\boxed{X = X_2}$**

**Se $Y = Y_3$ ⟶ $\boxed{X = X_3}$**

**Se $Y = Y_4$ ⟶ $\boxed{X = X_3}$**

# Perfetta Interdipendenza

**L'interdipendenza perfetta può essere raggiunta solo nel caso di tabella quadrata**

| | | Carattere Y | | | |
|---|---|---|---|---|---|
| | | $Y_1$ | $Y_2$ | $Y_3$ | Totale |
| Carattere X | $X_1$ | 25 | 0 | 0 | 25 |
| | $X_2$ | 0 | 0 | 30 | 30 |
| | $X_3$ | 0 | 45 | 0 | 45 |
| | Totale | 25 | 45 | 30 | 100 |

$X = X_1 \longleftrightarrow Y = Y_1$

$X = X_2 \longleftrightarrow Y = Y_3$

$X = X_3 \longleftrightarrow Y = Y_2$

**Interdipendenza (perfetta): ad ogni modalità del carattere X corrisponde una ed una sola modalità di Y e, simultaneamente, ad ogni modalità del carattere Y corrisponde una ed una sola modalità di X**

# Misurare la Dipendenza – il $\chi^2$

$$\chi^2 = \sum_{i=1}^{H} \sum_{j=1}^{K} \frac{\left(n_{ij} - n'_{ij}\right)^2}{n'_{ij}}$$

| | | | | |
|---|---|---|---|---|
| **Indipendenza** | → | $n_{ij} = n'_{ij}$ | → | $\chi^2 = 0$ |
| **Dipendenza** | → | $n_{ij} \neq n'_{ij}$ | → | $\chi^2 > 0$ |

**Massima dipendenza** → $\chi^2_{max} = n \cdot \min\left[(H-1);(K-1)\right]$

# Misurare la Dipendenza – il $\chi^2$

$$V = \sqrt{\frac{\chi^2}{\chi^2_{\max}}}$$

**V di Cramer**

**Tale indice può variare tra zero ed uno, e sarà pari a zero nel caso di indipendenza, mentre assumerà valore 1 nel caso di massima dipendenza.**

# Esempio

**Abbiamo effettuato un'indagine sugli studenti del nostro Ateneo al fine di rilevare il grado di associazione tra il voto da questi riportato nell'esame di matematica e nell'esame di statistica. I risultati sono riportati nella tabella seguente (in riga i voti di matematica, in colonna quelli di statistica):**

| | 18-22 | 23-26 | 27-30 | |
|---|---|---|---|---|
| 18-22 | 20 | 3 | 2 | 25 |
| 23-26 | 2 | 27 | 6 | 35 |
| 27-30 | 4 | 6 | 30 | 40 |
| | 26 | 36 | 38 | 100 |

**Si calcoli l'associazione tra queste due variabili utilizzando l'indice Chi-quadrato e la V di Cramer**

# Esempio - 2

### b) Frequenze teoriche

| | 18 – 22 | 23 – 26 | 27 – 30 | Tot |
|---|---|---|---|---|
| 18 – 22 | $\frac{25 \cdot 26}{100} = 6{,}5$ | $\frac{25 \cdot 36}{100} = 9$ | $\frac{25 \cdot 38}{100} = 9{,}5$ | 25 |
| 23 – 26 | $\frac{35 \cdot 26}{100} = 9{,}1$ | $\frac{35 \cdot 36}{100} = 12{,}6$ | $\frac{35 \cdot 38}{100} = 13{,}3$ | 35 |
| 27 – 30 | $\frac{40 \cdot 26}{100} = 10{,}4$ | $\frac{40 \cdot 36}{100} = 14{,}4$ | $\frac{40 \cdot 38}{100} = 15{,}2$ | 40 |
| Tot | 26 | 36 | 38 | 100 |

### c) Differenza (Frequenze effettive - Frequenze teoriche)

| | 18 – 22 | 23 – 26 | 27 – 30 | Tot |
|---|---|---|---|---|
| 18 – 22 | $(20 - 6{,}5) = 13{,}5$ | $(3 - 9) = -6$ | $(2 - 9{,}5) = -7{,}5$ | 0 |
| 23 – 26 | $(2 - 9{,}1) = -7{,}1$ | $(27 - 12{,}6) = 14{,}4$ | $(6 - 13{,}3) = -7{,}3$ | 0 |
| 27 – 30 | $(4 - 10{,}4) = -6{,}4$ | $(6 - 14{,}4) = -8{,}4$ | $(30 - 15{,}2) = 14{,}8$ | 0 |
| Tot | 0 | 0 | 0 | 0 |

# Esempio - 3

**d) Differenza (Frequenze effettive - Frequenze teoriche)$^2$**

| | 18 – 22 | 23 – 26 | 27 – 30 | Tot |
|---|---|---|---|---|
| 18 – 22 | $13{,}5^2 = 182{,}25$ | $-6^2 = 36$ | $-7{,}5^2 = 56{,}25$ | |
| 23 – 26 | $-7{,}1^2 = 50{,}41$ | $14{,}4^2 = 207{,}36$ | $-7{,}3^2 = 53{,}29$ | |
| 27 – 30 | $-6{,}4^2 = 40{,}96$ | $-8{,}4^2 = 70{,}56$ | $14{,}8^2 = 219{,}04$ | |
| Tot | | | | |

**e) Differenza (Frequenze effettive -Frequenze teoriche)$^2$diviso teoriche**

| | 18 – 22 | 23 – 26 | 27 – 30 | Tot |
|---|---|---|---|---|
| 18 – 22 | $\frac{182{,}25}{6{,}5} = 28{,}04$ | $\frac{36}{9} = 4$ | $\frac{56{,}25}{9{,}5} = 5{,}92$ | |
| 23 – 26 | $\frac{50{,}41}{9{,}1} = 5{,}54$ | $\frac{207{,}36}{12{,}6} = 16{,}46$ | $\frac{53{,}29}{13{,}3} = 4{,}01$ | |
| 27 – 30 | $\frac{40{,}96}{10{,}4} = 3{,}94$ | $\frac{70{,}56}{14{,}4} = 4{,}9$ | $\frac{219{,}04}{15{,}2} = 14{,}41$ | |
| Tot | | | | 87,21 |

# Esempio - 4

$$\chi^2 = \sum_{i=1}^{H}\sum_{j=1}^{K}\frac{\left(n_{ij} - n'_{ij}\right)^2}{n'_{ij}} = 87{,}21$$

$$\chi^2_{max} = n \cdot \min\left[(H-1);(K-1)\right] = 100 \cdot (3-1) = 100 \cdot 2 = 200$$

$$V = \sqrt{\frac{\chi^2}{\chi^2_{max}}} = \sqrt{\frac{87{,}21}{200}} = \sqrt{0{,}436} = 0{,}66$$

# Esercizio 2

**Ad un gruppo di individui che hanno contratto una certa malattia viene somministrata una medicina con differenti dosaggi (in mg). La condizione dei pazienti in seguito al trattamento è stata riportata nella tabella sottostante:**

| | 0-20 | 20-40 | 40-100 | |
|---|---|---|---|---|
| Peggiore | 18 | 4 | 2 | 24 |
| Invariata | 2 | 8 | 7 | 17 |
| Migliore | 0 | 6 | 28 | 34 |
| | 20 | 18 | 37 | 75 |

**Si calcoli l'associazione tra dosaggio e condizione del paziente, utilizzando l'indice Chi-quadrato e la V di Cramer.**

# Esercizio 2 - 2

**a) Frequenze effettive**

| | 0-20 | 20-40 | 40-100 | |
|---|---|---|---|---|
| Peggiore | 18 | 4 | 2 | 24 |
| Invariata | 2 | 8 | 7 | 17 |
| Migliore | 0 | 6 | 28 | 34 |
| | 20 | 18 | 37 | 75 |

**b) Frequenze teoriche**

| | 0-20 | 20-40 | 40-100 | |
|---|---|---|---|---|
| Peggiore | 6,40 | 5,76 | 11,84 | 24,00 |
| Invariata | 4,53 | 4,08 | 8,39 | 17,00 |
| Migliore | 9,07 | 8,16 | 16,77 | 34,00 |
| | 20,00 | 18,00 | 37,00 | 75,00 |

# Esercizio 2 - 3

**Differenza (Frequenze effettive -Frequenze teoriche)**

| | 0-20 | 20-40 | 40-100 | |
|---|---|---|---|---|
| Peggiori | 11,60 | -1,76 | -9,84 | 0,00 |
| Invariate | -2,53 | 3,92 | -1,39 | 0,00 |
| Migliori | -9,07 | -2,16 | 11,23 | 0,00 |
| | 0,00 | 0,00 | 0,00 | 0,00 |

**(18 - 6,40) = 11,60**

**(7 - 8,39) = -1,39**

**(6 - 8,16) = -2,16**

**Differenza (Frequenze effettive -Frequenze teoriche)$^2$**

| | 0-20 | 20-40 | 40-100 |
|---|---|---|---|
| Peggiori | 134,56 | 3,09 | 96,82 |
| Invariate | 6,40 | 15,36 | 1,93 |
| Migliori | 82,26 | 4,66 | 126,11 |

**$(11,60)^2$ = 134,56**

**$(3,92)^2$ = 15,36**

**$(-9,84)^2$ = 96,82**

# Esercizio 2 - 4

**Differenza (Frequenze effettive -Frequenze teoriche)$^2$ diviso teoriche**

| | 0-20 | 20-40 | 40-100 | |
|---|---|---|---|---|
| Peggiori | 21,02 | 0,53 | 8,17 | |
| Invariate | 1,41 | 3,76 | 0,23 | |
| Migliori | 9,06 | 0,57 | 7,51 | |
| | | | | 52,26 |

**(134,56 : 6,40) = 21,02**

**(1,93 : 8,39) = 0,23**

**Alta associazione**

$$\chi^2 = \sum_{i=1}^{H}\sum_{j=1}^{K}\frac{(n_{ij} - n'_{ij})^2}{n'_{ij}} = 52,26$$

$$\chi^2_{max} = n \cdot \min[(H-1);(K-1)] = 75 \cdot (3-1) = 75 \cdot 2 = 150$$

$$V = \sqrt{\frac{\chi^2}{\chi^2_{max}}} = \sqrt{\frac{52,26}{150}} = \sqrt{0,348} = 0,589$$

# Alcune considerazioni finali

**Condizioni paziente - Dosaggio**

| | 0-20 | 20-40 | 40-100 | |
|---|---|---|---|---|
| Peggiore | 18 | 4 | 2 | 24 |
| Invariata | 2 | 8 | 7 | 17 |
| Migliore | 0 | 6 | 28 | 34 |
| | 20 | 18 | 37 | 75 |

$$V = 0{,}589$$

**rischio – età del paziente**

| | 15-25 | 26-40 | 41-65 | |
|---|---|---|---|---|
| Moderato | 0 | 12 | 21 | 33 |
| Medio | 2 | 18 | 12 | 32 |
| Alto | 32 | 6 | 0 | 38 |
| | 34 | 36 | 33 | 103 |

$$\chi^2 = 78{,}87$$

$$V = 0{,}618$$

# Il test Chi-quadrato

**Ad un gruppo di individui che hanno contratto una certa malattia viene somministrata una medicina con differenti dosaggi (in mg). La condizione dei pazienti in seguito al trattamento è stata riportata nella tabella sottostante:**

| | 0-20 | 20-40 | 40-100 | |
|---|---|---|---|---|
| Peggiore | 18 | 4 | 2 | 24 |
| Invariata | 2 | 8 | 7 | 17 |
| Migliore | 0 | 6 | 28 | 34 |
| | 20 | 18 | 37 | 75 |

**Si calcoli l'associazione tra dosaggio e condizione del paziente, utilizzando l'indice Chi-quadrato e la V di Cramer.**

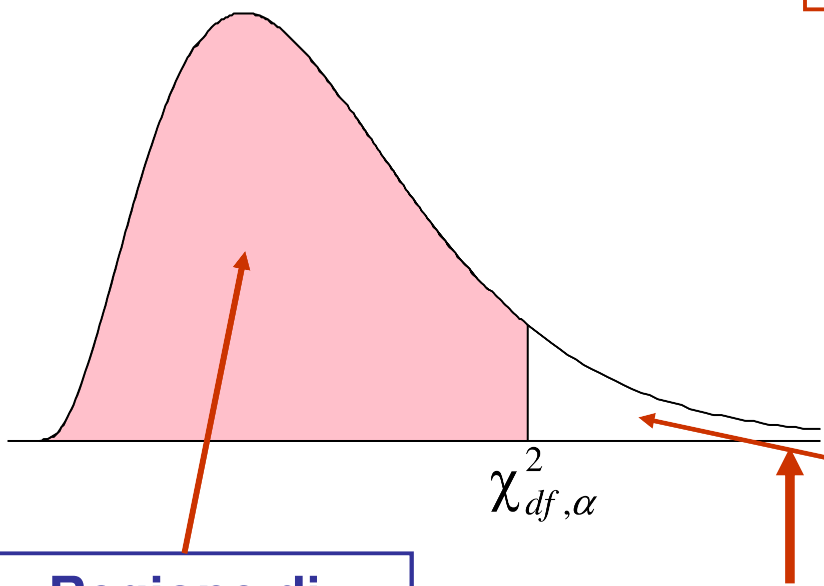
$$V = \sqrt{\frac{\chi^2}{\chi^2_{max}}} = \sqrt{\frac{52,26}{150}} = \sqrt{0,348} = 0,589$$

**Alta associazione**

# Il test Chi-quadrato - 2

$H_0 \rightarrow \quad \chi^2 = 0$ (**Indipendenza**)

$H_1 \rightarrow \quad \chi^2 \neq 0$ (**Dipendenza**)



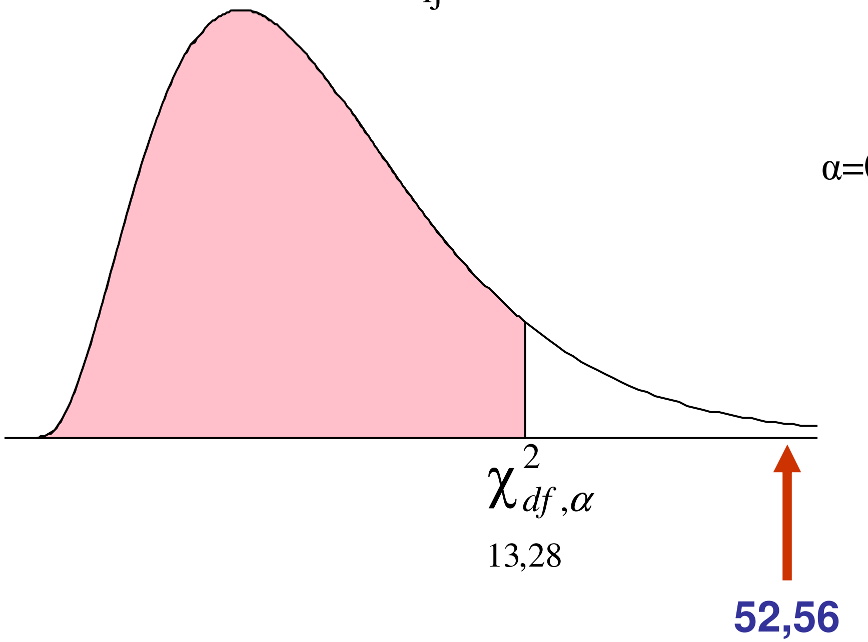

**Regione di** **<u>accettazione</u>:**
**$H_0$ Vera**

**52,26**

**Regione di** **<u>rifiuto</u>:**
**$H_0$ Falsa**

# Il test Chi-quadrato - 2

$$\chi^2 = \sum_{i=1}^{n} \frac{(n_{ij} - n'_{ij})^2}{n'_{ij}} = 52{,}56$$




**Rifiuto Ipotesi $H_0$**

**$H_0 \rightarrow \chi^2 = 0$ (Indipendenza)**

**$H_1 \rightarrow \chi^2 \neq 0$ (Dipendenza)**

α=0,01

$$df = (r-1)\cdot(c-1) = 2\times 2 = 4$$

$$\chi^2_{4;\,0,01} = 13{,}28$$

| gdl | 0,100 | 0,050 | 0,025 | 0,010 |
|---|---|---|---|---|
| 1 | 2,706 | 3,841 | 5,024 | 6,635 |
| 2 | 4,605 | 5,991 | 7,378 | 9,210 |
| 3 | 6,251 | 7,815 | 9,348 | 11,345 |
| 4 | 7,779 | 9,488 | 11,143 | 13,277 |
| 5 | 9,236 | 11,070 | 12,833 | 15,086 |
| 6 | 10,645 | 12,592 | 14,449 | 16,812 |
| 7 | 12,017 | 14,067 | 16,013 | 18,475 |
| 8 | 13,362 | 15,507 | 17,535 | 20,090 |
| 9 | 14,684 | 16,919 | 19,023 | 21,666 |
| 10 | 15,987 | 18,307 | 20,483 | 23,209 |

# Riferimenti sul testo

di **Whitlock M.C., Schluter D.**
***Analisi statistica dei dati biologici*, Zanichelli**

**Paragrafi da studiare: 9.1, 9.3. Esercizi alla fine dei paragrafi**