

Esercitazione 3/4

Corso di Statistica

Info

Daniele Pasquini

Phd Student in Data Science

- **Email:** daniele.pasquini@uniroma2.eu
- **Ricevimento:** Online su Teams
- **Ricevimento:** Ing. Informatica, Corpo D, Piano 2, Stanza 20

Concentrazione

Carattere **quantitativo trasferibile**:

- può essere ripartito tra le unità, come reddito, ricchezza, vendite, produzione
- carattere numerico che si può **sommare** e **dividere** fra le unità
- la **somma dei valori** ha senso economico o fisico

Concentrazione

Esempi :

- **Reddito individuale:** può essere sommato → reddito totale della popolazione.
- **Ricchezza posseduta:** si può sommare o distribuire tra persone.
- **Vendite di imprese:** somma delle vendite = fatturato totale del mercato.
- **Produzione agricola** di ogni azienda = totale della produzione nel settore.

Concentrazione

Non trasferibili:

- Età, altezza, peso, numero di figli, voto d'esame, ecc.
- Non si "sommano" in modo significativo per ottenere un totale collettivo.

Concentrazione

$x_1, x_2, \dots, x_n$ valori del carattere per le n unità statistiche

- **Totale complessivo:**

Concentrazione

$x_1, x_2, \dots, x_n$ valori del carattere per le n unità statistiche

- **Quota relativa** del carattere per ciascuna unità:

$$p_i = \frac{x_i}{A}$$

- Con le proprietà: $p_i \geq 0$ e $\sum_{i=1}^n p_i = 1$

Concentrazione

- **Caso limite:**

$$x_i = \frac{A}{n} \Rightarrow p_i = \frac{1}{n}$$

- **Equidistribuzione**

- Tutte le unità possiedono la stessa quota del carattere
- Ogni unità detiene la stessa parte del totale

Concentrazione

- **Caso limite:**

$$x_1 = A, \quad x_2 = x_3 = \dots = x_n = 0$$

- **Massima concentrazione**
- Tutto l'ammontare è posseduto da una sola unità
- Una sola unità possiede tutto $\Rightarrow$ concentrazione massima

Concentrazione

Persona	Reddito (€)	Quota $\pi_i = x_i/A$
1	20 000	0.10
2	20 000	0.10
3	20 000	0.10
4	20 000	0.10
5	120 000	0.60
Totale	200 000	1.00

Il grado di concentrazione misura **quanto la distribuzione reale** si discosta da quella di equidistribuzione.

Indice di Gini

Misura quanto la **distribuzione osservata** si discosta dall'equidistribuzione perfetta

- $G = 0 \rightarrow$ **equidistribuzione** perfetta (tutti possiedono la stessa quota)
- $G = 1 \rightarrow$ **massima concentrazione** (una sola unità possiede tutto)

Indice di Gini

- Ordinando i valori x_i in ordine crescente ($x_1 \leq x_2 \leq \dots \leq x_n$)

$$G = \frac{\sum_{i=1}^n \sum_{j=1}^n |x_i - x_j|}{2n^2 \bar{x}}$$

dove:

- x_i = valore del carattere per l'unità i
- $\left(\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \right)$ = media aritmetica dei valori
- n = numero totale di unità

Indice di Gini

- Ordinando i valori x_i in ordine crescente ($x_1 \leq x_2 \leq \dots \leq x_n$)

$$G = \frac{2 \sum_{i=1}^n i x_i}{n \sum_{i=1}^n x_i} - \frac{n+1}{n}$$

dove:

- i = è il rango (posizione ordinata) dell'unità
- (x_i) = valore del carattere per quella unità

Indice di Gini

Unità	x_i	Rango i
1	10	1
2	20	2
3	30	3

$$G = \frac{2(1 \cdot 10 + 2 \cdot 20 + 3 \cdot 30)}{3 \cdot 60} - \frac{3 + 1}{3}$$

$$G = \frac{2(140)}{180} - \frac{4}{3} = 1.555 - 1.333 = 0.222$$

$G = 0.22 \rightarrow$ bassa concentrazione, quindi
distribuzione abbastanza equa

Associazione tra 2 caratteri

- si osservano **due caratteri** statistici nello stesso insieme di unità
- esiste una **relazione** tra le modalità dei due caratteri?
- Una **distribuzione doppia** rappresenta la **frequenza congiunta** con cui si presentano simultaneamente due modalità, una per ciascun carattere

Associazione tra 2 caratteri

Si consideri una collettività di **n unità statistiche** e siano:

- **X**: primo carattere con modalità $x_1, x_2, \dots, x_r$
- **Y**: secondo carattere con modalità $y_1, y_2, \dots, y_s$

Associazione tra 2 caratteri

tabella a doppia entrata di frequenze congiunte

n_{ij}

Y \ X	x_1	x_2	...	x_r	Totale (Y)
y_1	n_{11}	n_{21}	...	n_{r1}	$n_{.1}$
y_2	n_{12}	n_{22}	...	n_{r2}	$n_{.2}$
...	...	...	...	...	...
y_s	n_{1s}	n_{2s}	...	n_{rs}	$n_{.s}$
Totale (X)	$n_{1.}$	$n_{2.}$	...	$n_{r.}$	n

Frequenze marginali

Le **distribuzioni marginali** rappresentano la distribuzione di ciascun carattere preso singolarmente

- Marginali di riga (per il carattere X) :

$$n_{i.} = \sum_{j=1}^s n_{ij}$$

Frequenze marginali

Le **distribuzioni marginali** rappresentano la distribuzione di ciascun carattere preso singolarmente

- Marginali di colonna (per il carattere Y) :

$$n_{.j} = \sum_{i=1}^r n_{ij}$$

Frequenze marginali

Le **distribuzioni marginali** rappresentano la distribuzione di ciascun carattere preso singolarmente

- Totale complessivo:

$$n = \sum_{i=1}^r \sum_{j=1}^s n_{ij}$$

Frequenze relative

Si ottengono dividendo ogni **frequenza assoluta** per il totale n :

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

$$f_{i.} = \frac{n_{i.}}{n} \quad ; \quad f_{.j} = \frac{n_{.j}}{n}$$

In questo modo:

$$\sum_{i=1}^r \sum_{j=1}^s f_{ij} = 1$$

Frequenze relative

Y \ X	x_1	x_2	Totale (Y)
y_1	20	10	30
y_2	30	40	70
Totale (X)	50	50	100

$f_{11} = 20/100 = 0.20$ frequenza relativa congiunta

$f_{1.} = 50/100 = 0.50$ frequenza relativa marginale

$f_{.1} = 30/100 = 0.30$ frequenza relativa marginale

Frequenze relative

Una classe di 50 studenti è stata classificata in base al **genere** e al **giudizio in statistica**

	Sufficiente	Buono	Ottimo	Totale
Maschi	10	8	7	25
Femmine	5	12	8	25
Totale	15	20	15	50

1. Calcola le frequenze relative congiunte
2. Calcola le frequenze relative marginali

Frequenze relative

Una classe di 50 studenti è stata classificata in base al **genere** e al **giudizio in statistica**

	Sufficiente	Buono	Ottimo	Totale
Maschi	10	8	7	25
Femmine	5	12	8	25
Totale	15	20	15	50

Riga Maschi

$$f_{11} = \frac{10}{50} = 0.20$$

$$f_{12} = \frac{8}{50} = 0.16$$

$$f_{13} = \frac{7}{50} = 0.14$$

Riga Femmine

$$f_{21} = \frac{5}{50} = 0.10$$

$$f_{22} = \frac{12}{50} = 0.24$$

$$f_{23} = \frac{8}{50} = 0.16$$

Marginali di riga

$$\text{Maschi: } \frac{25}{50} = 0.50$$

$$\text{Femmine: } \frac{25}{50} = 0.50$$

Marginali di colonna

$$\text{Sufficiente: } \frac{15}{50} = 0.30$$

$$\text{Buono: } \frac{20}{50} = 0.40$$

$$\text{Ottimo: } \frac{15}{50} = 0.30$$

Frequenze relative

Una classe di 50 studenti è stata classificata in base al **genere** e al **giudizio in statistica**

	Sufficiente	Buono	Ottimo	Totale
Maschi	10	8	7	25
Femmine	5	12	8	25
Totale	15	20	15	50

	Sufficiente	Buono	Ottimo	Totale
Maschi	0.20	0.16	0.14	0.50
Femmine	0.10	0.24	0.16	0.50
Totale	0.30	0.40	0.30	1.00

Distribuzioni condizionate

Mostrano come si distribuisce un carattere, fissata una modalità dell'altro carattere.

Distribuzione di **X** condizionata a **Y** = y_j :

$$f_{i|j} = \frac{f_{ij}}{f_{.j}} = \frac{n_{ij}}{n_{.j}}$$

Distribuzioni condizionate

Mostrano come si distribuisce un carattere, fissata una modalità dell'altro carattere.

- Distribuzione di **Y condizionata a $X = x_i$** :

$$f_{j|i} = \frac{f_{ij}}{f_{i.}} = \frac{n_{ij}}{n_{i.}}$$

Distribuzioni condizionate

Mostrano come si distribuisce un carattere, fissata una modalità dell'altro carattere

- Queste frequenze condizionate sommano a 1 per ogni valore fissato del carattere condizionante:

$$\sum_i f_{i|j=1} \quad \text{e} \quad \sum_j f_{j|i=1}$$

Distribuzioni condizionate

Una classe di 50 studenti è stata classificata in base al **genere** e al **giudizio in statistica**

	Sufficiente	Buono	Ottimo	Totale
Maschi	10	8	7	25
Femmine	5	12	8	25
Totale	15	20	15	50

1. Calcola la **distribuzione condizionata del giudizio dato il genere maschile**
2. Calcola la **distribuzione condizionata del genere dato il giudizio Buono**
3. quali combinazioni sono più probabili?

Distribuzioni condizionate

Una classe di 50 studenti è stata classificata in base al **genere** e al **giudizio in statistica**

	Sufficiente	Buono	Ottimo	Totale
Maschi	10	8	7	25
Femmine	5	12	8	25
Totale	15	20	15	50

Distribuzioni di genere dato il giudizio $f_{i|j} = \frac{n_{ij}}{n_{.j}}$

Sufficiente ($n_{.j} = 15$)

$$P(\text{Maschi} \mid \text{Suff}) = \frac{10}{15} \approx 0.6667$$

$$P(\text{Femmine} \mid \text{Suff}) = \frac{5}{15} \approx 0.3333$$

Buono ($n_{.j} = 20$)

$$P(\text{Maschi} \mid \text{Buono}) = \frac{8}{20} = 0.40$$

$$P(\text{Femmine} \mid \text{Buono}) = \frac{12}{20} = 0.60$$

Ottimo ($n_{.j} = 15$)

$$P(\text{Maschi} \mid \text{Ottimo}) = \frac{7}{15} \approx 0.4667$$

$$P(\text{Femmine} \mid \text{Ottimo}) = \frac{8}{15} \approx 0.5333$$

Indipendenza statistica

Due caratteri X e Y sono **statisticamente indipendenti** se la frequenza congiunta può essere espressa come prodotto delle marginali:

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

in termini assoluti:

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

Ovvero la distribuzione condizionata di un carattere non dipende dalla modalità dell'altro carattere

Indipendenza statistica

- Se $f_{ij} > f_{i.}f_{.j}$ → le modalità x_i e y_j tendono a comparire insieme più spesso del caso indipendente → **associazione positiva**
- Se $f_{ij} < f_{i.}f_{.j}$ → le modalità tendono a escludersi → **associazione negativa**

Indipendenza statistica

Y \ X	x1	x2	Totale (Y)
y1	20	10	30
y2	30	40	70
Totale (X)	50	50	100

Le frequenze marginali sono i totali per riga e per colonna (cioè la distribuzione di ciascun carattere preso da solo).

- Totali di colonna (carattere X) :

$$n_{1.} = 50, \quad n_{2.} = 50$$

- Totali di riga (carattere Y) :

$$n_{.1} = 30, \quad n_{.2} = 70$$

Indipendenza statistica

Y \ X	x1	x2	Totale (Y)
y1	20	10	30
y2	30	40	70
Totale (X)	50	50	100

<i>Y\X</i>	<i>f</i> _{1j}	<i>f</i> _{2j}	<i>Totale(Y)</i>
<i>y</i> ₁	$f_{11} = \frac{20}{100} = 0.20$	$f_{21} = \frac{10}{100} = 0.10$	$f_{.1} = 0.30$
<i>y</i> ₂	$f_{12} = \frac{30}{100} = 0.30$	$f_{22} = \frac{40}{100} = 0.40$	$f_{.2} = 0.70$
<i>Totale(X)</i>	$f_{1.} = 0.50$	$f_{2.} = 0.50$	1.00

Le frequenze relative si ottengono dividendo ogni frequenza assoluta per il totale (n = 100) .

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

Indipendenza statistica

Y \ X	x1	x2	Totale (Y)
y1	20	10	30
y2	30	40	70
Totale (X)	50	50	100

Calcolo teorico indipendenza

- Per la cella (x_1, y_1) :

$$f_{11}^{(teorico)} = f_{1.} \cdot f_{.1} = 0.50 \times 0.30 = 0.15$$

Indipendenza statistica

Y \ X	x1	x2	Totale (Y)
y1	20	10	30
y2	30	40	70
Totale (X)	50	50	100

- Frequenza osservata:

$$f_{11} = 0.20$$

- Frequenza attesa in caso di indipendenza:

$$f_{11}^{(teorico)} = 0.15$$

$$f_{11} = 0.20 > 0.15$$

significa che la combinazione $X = x_1$ e $Y = y_1$ **si verifica più spesso di quanto ci si aspetterebbe se X e Y fossero indipendenti**

Indipendenza

Una classe di 50 studenti è stata classificata in base al **genere** e al **giudizio in statistica**

	Sufficiente	Buono	Ottimo	Totale
Maschi	10	8	7	25
Femmine	5	12	8	25
Totale	15	20	15	50

1. Calcola le frequenze attese (indipendenza)

Indipendenza

Una classe di 50 studenti è stata classificata in base al **genere** e al **giudizio in statistica**

	Sufficiente	Buono	Ottimo	Totale
Maschi	10	8	7	25
Femmine	5	12	8	25
Totale	15	20	15	50

	Suff.	Buono	Ottimo	Totale
Maschi	7.5	10	7.5	25
Femmine	7.5	10	7.5	25
Totale	15	20	15	50

Indipendenza

Interpretazione preliminare

Cella	Osservata	Attesa	Confronto
Maschi-Suff.	10	7.5	Osservata più alta
Maschi-Buono	8	10	Osservata più bassa
Maschi-Ottimo	7	7.5	quasi uguale
Femmine-Suff.	5	7.5	più bassa
Femmine-Buono	12	10	più alta
Femmine-Ottimo	8	7.5	quasi uguale

Misure di associazione

Quando i caratteri X e Y sono **qualitativi** (cioè non numerici)...

- l'associazione si studia tramite una **tabella di contingenza**, formata dalle **frequenze congiunte** n_{ij}

Misure di associazione

- Immaginiamo di aver intervistato 60 persone e di aver ottenuto questa tabella di contingenza (frequenze assolute n_{ij})

	Reddito Basso (Y=1)	Reddito Medio (Y=2)	Reddito Alto (Y=3)	Totale riga
Studio Basso (X=1)	15	5	0	20
Studio Medio (X=2)	5	15	5	25
Studio Alto (X=3)	0	5	10	15
Totale colonna	20	25	15	60

Misure di associazione

Test di indipendenza – Chi quadrato di Pearson

- Serve per verificare se esiste **un'associazione** tra i due caratteri

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \text{ e } gl = (r - 1)(s - 1)$$

somma, cella per cella, **quanto gli osservati differiscono dagli attesi**, rapportato all'atteso

Misure di associazione

Test di indipendenza – Chi quadrato di Pearson

- Frequenze relative: $f_{ij} = \frac{n_{ij}}{n}$; $f_{i.} = \frac{n_{i.}}{n}$; $f_{.j} = \frac{n_{.j}}{n}$
- Se X e Y indipendenti $f_{ij} = f_{i.} \cdot f_{.j}$
- Moltiplichiamo entrambi i lati per n: $n_{ij} = n f_{ij} = n f_{i.} \cdot f_{.j}$
- Da cui si deriva che: $n_{ij} = \frac{n_{i.} \cdot n_{.j}}{n}$

Questa è la **formula delle frequenze attese**,
cioè quelle che ci aspettiamo se i due
caratteri fossero indipendenti.

Misure di associazione

Test di indipendenza – Chi quadrato di Pearson

Per verificare se i dati **seguono davvero** l'indipendenza, confrontiamo:

- le **frequenze osservate** n_{ij} ,
- con le **frequenze attese** $n_{ij}^{(teorico)}$

Se le differenze sono **piccole** → indipendenza plausibile.

Se sono **grandi** → probabilmente i due caratteri sono associati.

Misure di associazione

Test di indipendenza – Chi quadrato di Pearson

I gradi di libertà indicano **quante celle possono variare liberamente** prima che i totali di riga e colonna “blocchino” le altre

- Calcoli $\chi^2_{osservato}$

Confronti questo valore con quello **critico** della distribuzione Chi-quadrato, preso da una **tabella statistica** o da un software (es. Excel, R, Python, calcolatrice scientifica). Il valore critico dipende da:

- i **gradi di libertà** (gl),
- e il **livello di significatività** (es. $\alpha = 0.05$)

Se:

$\chi^2_{osservato} > \chi^2_{critico}$ si **rifiuta l'ipotesi di indipendenza**, quindi i due caratteri sono **associati**.

Test chi quadro

Una classe di 50 studenti è stata classificata in base al **genere** e al **giudizio in statistica**

	Sufficiente	Buono	Ottimo	Totale
Maschi	10	8	7	25
Femmine	5	12	8	25
Totale	15	20	15	50

1. Calcola le **frequenze attese** sotto l'ipotesi di indipendenza
2. Calcola la statistica chi quadro
3. Determina i **gradi di libertà**
4. Confronta il valore ottenuto con il valore critico (ad esempio a $\alpha = 0.05$)
5. i due caratteri (genere e giudizio) sono indipendenti?

Test chi quadro

Una classe di 50 studenti è stata classificata in base al **genere** e al **giudizio in statistica**

	Suff.	Buono	Ottimo	Totale
Maschi	7.5	10	7.5	25
Femmine	7.5	10	7.5	25
Totale	15	20	15	50

Contributi cella per cella:

$$\text{Maschi-Suff.: } (10 - 7.5)^2 / 7.5 = 0.8333$$

$$\text{Maschi-Buono: } (8 - 10)^2 / 10 = 0.4000$$

$$\text{Maschi-Ottimo: } (7 - 7.5)^2 / 7.5 = 0.0333$$

$$\text{Femmine-Suff.: } (5 - 7.5)^2 / 7.5 = 0.8333$$

$$\text{Femmine-Buono: } (12 - 10)^2 / 10 = 0.4000$$

$$\text{Femmine-Ottimo: } (8 - 7.5)^2 / 7.5 = 0.0333$$

$$\chi^2 \approx 0.8333 + 0.4000 + 0.0333 + 0.8333 + 0.4000 + 0.0333 = 2.533$$

$$gl = (r-1)(c-1) = (2-1)(3-1) = 2$$

per $gl = 2$ il valore critico a $\alpha = 0.05$ è circa 5.991; qui $2.533 < 5.991 \Rightarrow$ non significativo al 5%

Test chi quadro

Una classe di 50 studenti è stata classificata in base al **genere** e al **giudizio in statistica**

	Suff.	Buono	Ottimo	Totale
Maschi	7.5	10	7.5	25
Femmine	7.5	10	7.5	25
Totale	15	20	15	50

È importantissimo sottolineare:

- Il test **non dimostra** che i caratteri siano indipendenti.
- Dice solo che **non ci sono prove sufficienti** per affermare il contrario.

Misure di associazione

Coefficiente di contingenza di Pearson

- Misura normalizzata del χ^2 , che indica l'intensità dell'associazione (non la direzione)

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

Valori possibili:

- $C = 0 \rightarrow$ indipendenza perfetta
- C vicino al valore massimo (< 1) $\rightarrow$ forte associazione

Il valore massimo dipende da r e s , **quindi non confrontabile tra tabelle di dimensioni diverse**

Misure di associazione

Coefficiente di contingenza di Pearson

- Il valore massimo teorico di C non arriva mai a 1
- e dipende dal numero di righe e colonne della tabella.

Per esempio:

- In una tabella 2×2 , il massimo possibile di C è ~ 0.707
- In una tabella 3×3 , ~ 0.816
- In una tabella 4×4 , ~ 0.866

Misure di associazione

Coefficiente di contingenza di Pearson

Tabella	χ^2	n	$C = \sqrt{(\chi^2 / (\chi^2 + n))}$	Dimensioni
A	20	100	0.408	2×2
B	20	100	0.408	3×3

I valori numerici sembrano uguali, ma **nelle due tabelle il massimo teorico è diverso.**

Quindi non possiamo dire che l'associazione "è uguale" in forza nei due casi.

Misure di associazione

V di Cramer

- Misura derivata dal χ^2 corretta per la dimensione della tabella, quindi confrontabile

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(r - 1, s - 1)}}$$

Valori possibili:

- $V = 0 \rightarrow$ indipendenza
- $V = 1 \rightarrow$ associazione perfetta

È la misura più usata per caratteri qualitativi nominali

Cramer

Una classe di 50 studenti è stata classificata in base al **genere** e al **giudizio in statistica**

	Sufficiente	Buono	Ottimo	Totale
Maschi	10	8	7	25
Femmine	5	12	8	25
Totale	15	20	15	50

- 1. Calcola il coefficiente di Cramer
- 2. Interpreta il risultato

Cramer

Una classe di 50 studenti è stata classificata in base al **genere** e al **giudizio in statistica**

	Sufficiente	Buono	Ottimo	Totale
Maschi	10	8	7	25
Femmine	5	12	8	25
Totale	15	20	15	50

$$n = 50, \quad V = \sqrt{\frac{2.533}{50 \cdot 1}} = \sqrt{0.05066} \approx 0.225$$

Misure di associazione

Gamma di Goodman e Kruskal

- Misura la forza e la direzione dell'associazione tra due caratteri ordinali

Due coppie $(x_i, y_i), (x_j, y_j)$ sono:

- concordanti se $(x_i - x_j)(y_i - y_j) > 0$
- discordanti se $(x_i - x_j)(y_i - y_j) < 0$

Misure di associazione

Gamma di Goodman e Kruskal

- Misura la forza e la direzione dell'associazione tra due caratteri ordinali

Siano:

C = numero di coppie concordanti

D = numero di coppie discordanti

$$\gamma = \frac{C - D}{C + D}$$

$\gamma = 1 \rightarrow$ perfetta concordanza (associazione positiva perfetta)

$\gamma = -1 \rightarrow$ perfetta discordanza (associazione negativa perfetta)

$\gamma = 0 \rightarrow$ indipendenza (nessuna associazione)

Misure di associazione

Rho di Spearman

- Misura la correlazione tra ranghi delle due variabili (anch'essa per dati ordinali)

Assegnando i ranghi R_i e S_i dei caratteri X e Y:

$$\rho_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

$d_i = R_i - S_i \rightarrow$ differenza tra I ranghi

$n \rightarrow$ numero di unità

Misure di associazione

Rho di Spearman

- Misura la correlazione tra ranghi delle due variabili (anch'essa per dati ordinali)

Assegnando i ranghi R_i e S_i dei caratteri X e Y:

$$\rho_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

$\rho_s = 1 \rightarrow$ correlazione positiva perfetta (ranghi identici)

$\rho_s = -1 \rightarrow$ correlazione negativa perfetta (ranghi inversi)

$\rho_s = 0 \rightarrow$ nessuna correlazione tra ranghi

Misure di associazione

Rho di Spearman

- Dall'esempio... si calcola il reddito medio ponderato per ogni livello di studio:

$$\bar{y}_i = \frac{\sum_j y_j n_{ij}}{n_{i.}}$$

- $y_j = 1,2,3$ sono i livelli di reddito,
- n_{ij} sono le frequenze osservate nella tabella,
- $n_{i.}$ è il totale di riga (cioè quante persone hanno quel livello di studio)

Misure di associazione

$$\bar{y}_i = \frac{\sum_j y_j n_{ij}}{n_{i.}}$$

$y_j = 1,2,3$ sono i livelli di reddito,
 n_{ij} sono le frequenze osservate nella
tabella,
 $n_{i.}$ è il totale di riga (cioè quante persone
hanno quel livello di studio)

si calcola il reddito medio ponderato per
ogni livello di studio

	Reddito Basso (Y=1)	Reddito Medio (Y=2)	Reddito Alto (Y=3)	Totale riga
Studio Basso (X=1)	15	5	0	20
Studio Medio (X=2)	5	15	5	25
Studio Alto (X=3)	0	5	10	15
Totale colonna	20	25	15	60

Misure di associazione

Rho di Spearman

- Esempio per “studio basso”:

$$\overline{y_1} = \frac{1 \cdot 15 + 2 \cdot 5 + 3 \cdot 0}{20} = 1.25$$

Significa che chi ha “studio basso” ha reddito medio pari a 1.25, cioè prevalentemente basso

Misure di associazione

Rho di Spearman

Assegniamo i ranghi da 1 a 3 alle variabili:

Livello di Studio (X)	Rango R_x	Reddito Medio y_i	Rango R_y
Basso (1)	1	1.25	1
Medio (2)	2	2.00	2
Alto (3)	3	2.67	3

Misure di associazione

Rho di Spearman

Calcolo delle differenza tra ranghi:

i	R_x	R_y	$d_i = R_x - R_y$	d_i^2
1	1	1	0	0
2	2	2	0	0
3	3	3	0	0

$$\rho_s = 1$$

All'aumentare del livello di studio,
aumenta anche il reddito **in modo monotono**

Misure di associazione

Rho di Spearman

- Spearman non guarda quanto è grande la differenza, ma in che direzione cambia:
 - se X aumenta, anche Y aumenta?
 - oppure Y diminuisce?
- Quindi basta sapere l'ordine dei valori medi:

$$\overline{y_1} = 1.25 < \overline{y_2} = 2.00 < \overline{y_3} = 2.67$$

- All'aumentare del livello di studio (da 1 a 3), aumenta anche il reddito medio.
- Per Spearman, questo corrisponde a una **correlazione monotona crescente perfetta**, cioè $\rho_s = 1$

Misure di associazione

- χ^2 dice "c'è dipendenza, sì o no?"
- **V di Cramer** dice "quanto è forte la dipendenza?"
- **Gamma, Spearman** dicono "la relazione è monotona? cresce o decresce?"