

UNIVERSITÀ · STATISTICA DESCRITTIVA BIVARIATA

Indipendenza **stocastica** e indipendenza **in media**

Formulario operativo: definizioni, formule e **metodi pratici** per riconoscere il tipo di legame fra due caratteri a partire da una tabella a doppia entrata. Con esempi svolti e verificati.

parte 1

Indip. stocastica

Confronto fra distribuzioni;
teorema di fattorizzazione;
connessione e indice χ^2 .

parte 2

Indip. in media

Medie condizionate;
scomposizione della varianza;
rapporto di correlazione η^2 .

parte 3

Quadro & decisioni

Relazioni fra i due concetti, casi
limite e schema decisionale.

Tabella a doppia entrata e notazione



Due caratteri rilevati sulle stesse n unità si organizzano in una **tabella a doppia entrata**: X con h modalità sulle righe, Y con k modalità sulle colonne. Tutto ciò che segue si legge da qui.

n_{ij}	y_1	y_2	...	y_k	tot.
x_1	n_{11}	n_{12}	...	n_{1k}	$n_{1.}$
x_2	n_{21}	n_{22}	...	n_{2k}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_h	n_{h1}	n_{h2}	...	n_{hk}	$n_{h.}$
tot.	$n_{.1}$	$n_{.2}$	...	$n_{.k}$	n

Frequenze congiunte n_{ij} , marginali di riga $n_{i.}$ e di colonna $n_{.j}$.

SIMBOLI

- n_{ij} — frequenza **congiunta** (cella riga i , colonna j)
- $n_{i.} = \sum_j n_{ij}$ — marginale di riga
- $n_{.j} = \sum_i n_{ij}$ — marginale di colonna
- $n = \sum_i \sum_j n_{ij}$ — totale

Distribuzioni che useremo

FREQUENZE RELATIVE

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

$f_{i.}$ e $f_{.j}$ sono le distribuzioni **marginali** relative di X e di Y .

DISTRIBUZIONI CONDIZIONATE (RELATIVE)

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

$f_{j|i}$: distribuzione di Y **dentro** la riga x_i (ogni cella divisa per il totale di riga). Idea: «come si distribuisce Y quando fisso $X = x_i$ ».

IL FILO CONDUTTORE

Entrambe le forme di indipendenza rispondono alla stessa domanda — « **fissare una modalità di X cambia qualcosa su Y ?**» — ma a livelli diversi: l'indipendenza **stocastica** guarda l'intera distribuzione, l'indipendenza **in media** guarda solo la media.



Definizione e teorema di fattorizzazione



DEFINIZIONE

X e Y sono **indipendenti stocasticamente** quando tutte le distribuzioni condizionate relative sono uguali fra loro e uguali alla marginale: cambiare la modalità su cui si condiziona **non modifica** la distribuzione dell'altro carattere.

$$f_{j|i} = f_{.j} \quad \forall i, j \quad \Longleftrightarrow \quad f_{i|j} = f_{i.} \quad \forall i, j$$

TEOREMA DI FATTORIZZAZIONE · LA FORMA OPERATIVA

È la versione che si usa per i conti. X e Y sono indipendenti se e solo se le frequenze osservate coincidono con le **frequenze teoriche** $\hat{n}_{ij}$, per ogni cella:

$$n_{ij} = \hat{n}_{ij} = \frac{n_{i.} \cdot n_{.j}}{n} \quad \forall i, j \quad \Longleftrightarrow \quad f_{ij} = f_{i.} \cdot f_{.j} \quad \forall i, j$$

n_{ij} osservate $\hat{n}_{ij}$ teoriche di indipendenza basta **una** cella con $n_{ij} \neq \hat{n}_{ij}$ per escludere l'indipendenza

I due metodi pratici per una tabella

METODO A · DISTRIBUZIONI CONDIZIONATE

- Dividi ogni colonna (o riga) per il suo totale → ottieni le condizionate relative.
- Confrontale fra loro: **tutte identiche** → indipendenza.
- Più intuitivo, ma richiede di costruire tutte le colonne.

METODO B · FREQUENZE TEORICHE

- Calcola $\hat{n}_{ij} = n_{i.} \cdot n_{.j} / n$ cella per cella.
- Confronta con le osservate: **tutte uguali** → indipendenza.
- Più rapido e si collega direttamente a χ^2 (parte connessione).

PROPRIETÀ CHIAVE

L'indipendenza stocastica è **simmetrica**: X indep. da $Y \Longleftrightarrow Y$ indep. da X . Vale qualunque sia il tipo dei caratteri (qualitativi o quantitativi).



Scorciatoie per decidere al volo



Spesso si può **escludere** l'indipendenza senza completare i calcoli. Queste quattro regole fanno risparmiare tempo all'esame.

1 · UNO ZERO «TRADISCE»

Se c'è anche **un solo** $n_{ij} = 0$ ma le marginali $n_{i.}, n_{.j} > 0$, allora $\hat{n}_{ij} > 0 \neq 0$: **NON** indipendenti.

2 · TEORICA NON INTERA

Se qualche $\hat{n}_{ij} = n_{i.}n_{.j}/n$ **non è intero**, non può coincidere con un'osservata (sempre intera): **NON** indipendenti. Scarti subito.

3 · RIGHE PROPORZIONALI

Indipendenza $\iff$ le righe (o le colonne) congiunte sono **proporzionali** fra loro: ogni riga è multipla di un'altra, modalità per modalità.

4 · SIMMETRIA

Controlli un solo verso ($X \rightarrow Y$ oppure $Y \rightarrow X$): la conclusione vale automaticamente anche nell'altro.

ESERCIZIO SVOLTO · PROFESSIONE PATERNA (X) E SCOLARITÀ DEI FIGLI (Y)

$X \backslash Y$	media	diploma	laurea	$n_{i.}$
operaio	3	2	5	10
impieg.	9	6	15	30
dirig.	18	12	30	60
$n_{.j}$	30	20	50	100

Le righe sono proporzionali: riga 2 = $3 \times$ riga 1, riga 3 = $6 \times$ riga 1.

Verifica (metodo B). $\hat{n}_{11} = \frac{10 \cdot 30}{100} = 3$,
 $\hat{n}_{12} = \frac{10 \cdot 20}{100} = 2$, $\hat{n}_{13} = \frac{10 \cdot 50}{100} = 5 \dots$
 ogni $\hat{n}_{ij}$ coincide con l'osservata.

Condizionate relative $Y \mid x_i$ tutte uguali:

(0,3; 0,2; 0,5) per operaio, impiegato, dirigente

Conclusione: $n_{ij} = \hat{n}_{ij} \forall i, j \Rightarrow$ indipendenza stocastica.



Quanto si è lontani dall'indipendenza



Se non c'è indipendenza, c'è **connessione**. La si misura confrontando frequenze osservate e teoriche.

CONTINGENZE

$$c_{ij} = n_{ij} - \hat{n}_{ij}$$

Somme di riga, di colonna e totale **sempre nulle**.

Tutte le $c_{ij} = 0 \iff$ indipendenza.

INDICE χ^2 E VERSIONE NORMALIZZATA

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} = \sum_i \sum_j \frac{c_{ij}^2}{\hat{n}_{ij}}$$

$$\chi_N^2 = \frac{\chi^2}{n \cdot \min[(h-1), (k-1)]}, \quad 0 \leq \chi_N^2 \leq 1$$

CASI LIMITE

- $\chi_N^2 = 0 \rightarrow$ indipendenza stocastica (osservate = teoriche).
- $\chi_N^2 = 1 \rightarrow$ dipendenza funzionale (massima connessione).
- χ_N^2 è **simmetrico**:
 $\chi_N^2(X | Y) = \chi_N^2(Y | X)$.

ATTENZIONE · ASSOCIAZIONE SPURIA

$\chi_N^2 > 0$ non implica un legame **diretto**: una terza variabile **Z** può indurre associazione fra **X** e **Y**.
Condizionando su **Z**, l'associazione può sparire.

ESERCIZIO SVOLTO · METEO (X) E MEZZO DI TRASPORTO (Y) — $n = 100$

$X \backslash Y$	bici	bus	piedi	auto	$n_{i.}$
sereno	20	5	25	0	50
variab.	10	5	5	10	30
pioggia	0	10	0	10	20
$n_{.j}$	30	20	30	20	100

Presenza di zeri $\Rightarrow$ già si sa: non indipendenti.

Conti. $h = 3, k = 4 \Rightarrow \min[(h-1), (k-1)] = 2$.

$$\chi^2 = \frac{5^2}{15} + \frac{(-5)^2}{10} + \dots + \frac{6^2}{4} = 55,55$$

indice normalizzato

$$\chi_N^2 = \frac{55,55}{100 \cdot 2} = 0,28$$

Lettura: connessione bassa-moderata, lontana sia da 0 che da 1.



Medie condizionate e scomposizione della varianza



PREREQUISITO

Serve **almeno un carattere quantitativo** (qui Y). Non si lavora più su tutta la distribuzione, ma solo sulle **medie** (e varianze) condizionate.

MEDIE E VARIANZE CONDIZIONATE

$$\mu_Y(x_i) = \frac{1}{n_{i.}} \sum_j y_j n_{ij}$$

$$\sigma_Y^2(x_i) = \frac{1}{n_{i.}} \sum_j (y_j - \mu_Y(x_i))^2 n_{ij}$$

SCOMPOSIZIONE DELLA VARIANZA

$$\sigma_Y^2 = \underbrace{\sigma_B^2}_{\text{between}} + \underbrace{\sigma_W^2}_{\text{within}}$$

$\sigma_B^2 = V(\mu_Y(x_i))$ – varianza delle medie

$\sigma_W^2 = M(\sigma_Y^2(x_i))$ – media delle varianze

DEFINIZIONE

Y è **indipendente in media** da X quando le medie condizionate sono tutte uguali fra loro (e quindi uguali alla media marginale di Y):

$$\mu_Y(x_i) = \mu_Y \quad \forall i \quad \implies \quad \sigma_B^2 = V(\mu_Y(x_i)) = 0$$

RAPPORTO DI CORRELAZIONE · MISURA LA DIPENDENZA IN MEDIA

$$\eta_{Y|X}^2 = \frac{\sigma_B^2}{\sigma_Y^2} = 1 - \frac{\sigma_W^2}{\sigma_Y^2}, \quad 0 \leq \eta_{Y|X}^2 \leq 1$$

$\eta^2=0 \rightarrow$ indipendenza in media di Y da X $\eta^2=1 \rightarrow$ dipendenza funzionale di Y da X



Metodo pratico ed esempio svolto



COME SI PROCEDE SU UNA TABELLA (Y QUANTITATIVA)

- Calcola le medie condizionate $\mu_Y(x_i)$, una per ogni modalità di X (per ogni colonna/gruppo).
- **Tutte uguali** $\rightarrow Y$ indep. in media da $X \rightarrow \sigma_B^2 = 0 \rightarrow \eta_{Y|X}^2 = 0$. Fine.
- **Diverse** $\rightarrow$ c'è dipendenza in media: calcola σ_B^2, σ_W^2 e quindi $\eta_{Y|X}^2$.

ESERCIZIO SVOLTO · $Y \in \{1, 2, 3\}$ SU $X = x_1, \dots, x_4 - n = 100$

$Y \backslash X$	x_1	x_2	x_3	x_4	$n_{.j}$
1	10	0	40	5	55
2	0	30	0	0	30
3	10	0	0	5	15
$n_{i.}$	20	30	40	10	100

gruppo	$\mu_Y(x_i)$	$\sigma_Y^2(x_i)$
x_1	2	1
x_2	2	0
x_3	1	0
x_4	2	1

Le medie condizionate **non** sono tutte uguali ($\mu_Y(x_3) = 1$): c'è dipendenza in media.

$$\mu_Y = 1,6 \quad \sigma_B^2 = 2,8 - 1,6^2 = 0,24$$

$$\sigma_W^2 = \frac{30}{100} = 0,30 \quad \Rightarrow \quad \sigma_Y^2 = 0,54$$

$$\eta_{Y|X}^2 = \frac{0,24}{0,54} = 1 - \frac{0,30}{0,54} = 0,44$$

Lettura: la dipendenza in media è circa il 44%.

DIFFERENZA CRUCIALE DALL'INDIP. STOCASTICA

L'indipendenza in media **NON** è simmetrica: in generale $\eta_{Y|X}^2 \neq \eta_{X|Y}^2$. Y può essere indipendente in media da X mentre X non lo è da Y . (Per fare $\eta_{X|Y}^2$ serve che anche X sia quantitativa.)



Quadro di sintesi e schema decisionale



Cosa vale in ogni situazione

Situazione	Freq. condiz. relative	Medie condiz. $\mu_Y(x_i)$	χ_N^2	$\eta_{Y X}^2$
Indipendenza stocastica	tutte uguali	tutte uguali	0	0
Indip. in media (ma non stocastica)	diverse	tutte uguali	> 0	0
Connessione / dipendenza parziale	diverse	diverse	tra 0 e 1	tra 0 e 1
Dipendenza funzionale di Y da X	—	diverse, $\sigma_Y^2(x_i) = 0$	1 ($\geq$ un verso)	1

RELAZIONI DA RICORDARE

- Indip. stocastica $\Rightarrow$ indep. in media (sia $Y | X$ sia $X | Y$). **Non** vale il viceversa.
- Dip. funzionale di Y da $X \Rightarrow \eta_{Y|X}^2 = 1$; ma $\eta_{X|Y}^2$ può essere < 1 (tabella rettangolare, non biunivoca).
- χ_N^2 simmetrico; η^2 **non** simmetrico.

SCHEMA DECISIONALE · DATA UNA TABELLA

- **Entrambi qualitativi?** $\rightarrow$ solo stocastica/connessione. Vai al metodo $\hat{n}_{ij}$.
- **Almeno uno quantitativo?** $\rightarrow$ puoi fare anche l'indip. in media.
- **Stocastica:** cerca zeri o $\hat{n}_{ij}$; non interi $\rightarrow$ scarti subito; altrimenti confronta tutte le celle, poi χ_N^2 .
- **In media (Y quant.):** calcola $\mu_Y(x_i)$; tutte uguali $\rightarrow$ indep.; altrimenti $\eta_{Y|X}^2$.

IN UNA RIGA

Stocastica = «la distribuzione non cambia» (forte, simmetrica). **In media** = «la media non cambia» (più debole, direzionale). La prima implica la seconda, non il contrario.

