Marco Minozzo

Laurea in Economia del Commercio Internazionale
Laurea in Economia e Amministrazione delle Imprese
Università degli Studi di Verona (sede di Vicenza)

Calcolo delle Probabilità
Paradosso di Simpson

Il paradosso di Simpson (Simpson, 1951) può sorgere quando si voglia fornire al concetto di dipendenza condizionata un'interpretazione di tipo causale.

Si consideri il caso di semplici eventi. In questo caso può aversi che, dati tre eventi A, B e C, l'indipendenza condizionata di A e B, sia condizionatamente a C che a $\overline{C}$, non sia accompagnata dall'indipendenza incondizionata di A e B. Cioè, in generale,

$$\mathsf{P}(A \cap B|C) = \mathsf{P}(A|C) \cdot \mathsf{P}(B|C), \qquad (A \perp\!\!\!\perp B \mid C),$$

e

$$\mathsf{P}(A \cap B|\overline{C}) = \mathsf{P}(A|\overline{C}) \cdot \mathsf{P}(B|\overline{C}), \qquad (A \perp\!\!\!\perp B \mid \overline{C}),$$

non implicano che

$$\mathsf{P}(A \cap B) = \mathsf{P}(A) \cdot \mathsf{P}(B), \qquad (A \perp\!\!\!\perp B).$$

Esempio. Si considerino due mazzi di carte M_1 e M_2. Il mazzo M_1 sia un usuale mazzo da 52 carte composto dalle carte '1', '2', '3', '4', '5', '6', '7', '8', '9', '10', 'fante', 'donna' e 're', secondo i quattro semi 'cuori', 'quadri', 'fiori' e 'picche'. Il mazzo M_2 sia invece un mazzo ridotto di sole 20 carte composto dalle carte '1', '2', '3', '4', '5', '6', '7', '8', '9' e '10', secondo i due semi 'cuori' e 'quadri'.

Dopo aver adeguatamente mescolato tra loro i due mazzi, si consideri di estrarre a sorte una carta dal mazzo così ottenuto. Indicando con A l'evento che la carta estratta sia un '3' e con B l'evento che la carta estratta sia un 'cuori', ed indicando con C l'evento che la carta provenga dal primo mazzo, e con $\overline{C}$ l'evento che la carta estratta provenga dal secondo mazzo, si verifica che $A \perp\!\!\!\perp B \mid C$ e che $A \perp\!\!\!\perp B \mid \overline{C}$, ma non che $A \perp\!\!\!\perp B$. Infatti, considerando che i due mazzi sono costituiti come nelle tabelle seguenti

Mazzo M_1

	$\heartsuit$	$\overline{\heartsuit}$	
'3'	1	3	4
$\overline{\text{'3'}}$	12	36	48
	13	39	52

Mazzo M_2

	$\heartsuit$	$\overline{\heartsuit}$	
'3'	1	1	2
$\overline{\text{'3'}}$	9	9	18
	10	10	20

Mazzo $M_1 + M_2$

	$\heartsuit$	$\overline{\heartsuit}$	
'3'	2	4	6
$\overline{\text{'3'}}$	21	45	66
	23	49	72

è facile verificare che $\mathsf{P}(\text{'3'} \cap \text{'cuori'}|M_1) = 1/52$ è pari al prodotto di $\mathsf{P}(\text{'3'}|M_1) = 4/52$ per $\mathsf{P}(\text{'cuori'}|M_1) = 13/52$, e che $\mathsf{P}(\text{'3'} \cap \text{'cuori'}|M_2) = 1/20$ è pari al prodotto di $\mathsf{P}(\text{'3'}|M_2) = 2/20$ per $\mathsf{P}(\text{'cuori'}|M_2) = 10/20$. Mentre invece $\mathsf{P}(\text{'3'} \cap \text{'cuori'}) = 2/72$ non è pari al prodotto di $\mathsf{P}(\text{'3'}) = 6/72$ per $\mathsf{P}(\text{'cuori'}) = 23/72$.

Nel seguente esempio si mostra come, per tre eventi T, G e M, la mancanza di indipendenza tra i primi due, condizionatamente al terzo, non implica la mancanza di indipendenza tra i primi due marginalmente, cioè

$$T \not\perp\!\!\!\perp G \mid M \quad \text{e} \quad T \not\perp\!\!\!\perp G \mid \overline{M} \qquad \not\Rightarrow \qquad T \not\perp\!\!\!\perp G.$$

Esempio. Si consideri una popolazione di 76 pazienti costituita da 43 individui maschi e da 33 individui femmine, per cui considerando di estrarre un individuo a caso da questa popolazione la probabilità che questo sia maschio è data da $\mathsf{P}(\text{Maschio}) = 43/76$, e la probabilità che questo sia femmina è data da $\mathsf{P}(\text{Femmina}) = 33/76$. Si consideri anche che in questa popolazione una parte dei pazienti è stata sottoposta ad un certo trattamento T, mentre la parte rimanente è stata sottoposta ad un diverso trattamento $\overline{T}$. Inoltre, mentre alcuni pazienti sono guariti (G), altri non hanno raggiunto la guarigione ($\overline{G}$). Le tre tabelle seguenti riportano i 76 pazienti classificati secondo le tre variabili: ‘trattamento’, ‘guarigione’ e ‘sesso’.

Maschio	G	$\overline{\text{G}}$		Femmina	G	$\overline{\text{G}}$		Totale	G	$\overline{\text{G}}$	
T	10	5	15	T	10	15	25	T	20	20	40
$\overline{\text{T}}$	16	12	28	$\overline{\text{T}}$	2	6	8	$\overline{\text{T}}$	18	18	36
	26	17	43		12	21	33		38	38	76

Chiaramente, le probabilità congiunte relative alla terza tabella possono ottenersi dalle probabilità condizionate relative alle prime due tabelle secondo la formula delle probabilità totali, considerando che $M \cup F = \Omega$, e $M \cap F = \emptyset$. Ad esempio, la probabilità congiunta relativa alla prima cella della terza tabella è data da

$$\begin{aligned}\mathsf{P}(T \cap G) &= \mathsf{P}(T \cap G \cap M) + \mathsf{P}(T \cap G \cap F) \\ &= \mathsf{P}(T \cap G|M)\mathsf{P}(M) + \mathsf{P}(T \cap G|F)\mathsf{P}(F) \\ &= \frac{10}{43} \cdot \frac{43}{76} + \frac{10}{33} \cdot \frac{33}{76} = \frac{20}{76}.\end{aligned}$$

Per ognuna delle tre tabelle, le probabilità condizionate per riga sono date da

Maschio	G	$\overline{\text{G}}$		Femmina	G	$\overline{\text{G}}$		Totale	G	$\overline{\text{G}}$
T	0,667	0,333		T	0,400	0,600		T	0,500	0,500
$\overline{\text{T}}$	0,571	0,429		$\overline{\text{T}}$	0,250	0,750		$\overline{\text{T}}$	0,500	0,500

Perciò, mentre si ha che

$$T \not\perp\!\!\!\perp G \mid M, \qquad T \not\perp\!\!\!\perp G \mid F,$$

ed in particolare che il trattamento sembra avere un effetto positivo sulla guarigione nella sottopopolazione dei soli maschi, come nella sottopopolazione delle sole femmine, si ha che

$$T \perp\!\!\!\perp G,$$

cioè il trattamento sembra non avere più effetto quando si considera la popolazione nel suo complesso.

Nel seguente esempio si mostra come, per tre eventi T, G ed M, l’indipendenza tra i primi due, condizionatamente al terzo, non implica l’indipendenza tra i primi due marginalmente, cioè

$$T \perp\!\!\!\perp G \mid M \quad \text{e} \quad T \perp\!\!\!\perp G \mid \overline{M} \qquad \not\Rightarrow \qquad T \perp\!\!\!\perp G.$$

Esempio. Si consideri una popolazione di pazienti dove gli eventi T, $\overline{T}$, G, $\overline{G}$, M, $\overline{M} \equiv F$ sono definiti similmente all'esempio precedente. La popolazione è però ora costituita da 30 individui di cui 12 sono maschi e 18 sono femmine. Le tre tabelle seguenti riportano la struttura di tale popolazione sia in base al sesso che globalmente.

Maschio

	G	$\overline{\text{G}}$	
T	2	4	6
$\overline{\text{T}}$	2	4	6
	4	8	12

Femmina

	G	$\overline{\text{G}}$	
T	10	5	15
$\overline{\text{T}}$	2	1	3
	12	6	18

Totale

	G	$\overline{\text{G}}$	
T	12	9	21
$\overline{\text{T}}$	4	5	9
	16	14	30

Le probabilità condizionate per riga sono quindi date da

Maschio

	G	$\overline{\text{G}}$	
T	0,333	0,667	
$\overline{\text{T}}$	0,333	0,667	

Femmina

	G	$\overline{\text{G}}$	
T	0,667	0,333	
$\overline{\text{T}}$	0,667	0,333	

Totale

	G	$\overline{\text{G}}$	
T	0,571	0,429	
$\overline{\text{T}}$	0,444	0,556	

Si vede perciò che, mentre

$$T \perp\!\!\!\perp G \mid M, \qquad\qquad T \perp\!\!\!\perp G \mid F,$$

cioè che il trattamento non sembra avere un effetto positivo sulla guarigione nelle due sottopopolazioni, di maschi e di femmine, considerate singolarmente, marginalmente si ha che

$$T \not\perp\!\!\!\perp G,$$

ed in particolare che il trattamento sembra avere effetto positivo sulla guarigione nella popolazione considerata nel suo complesso.