Distribuzione di frequenza doppia

Si prendano in considerazione due caratteri: X con modalità x_i (i=1,2,…,h) e Y con modalità y_j (j=1,2,..,k).

La distribuzione doppia di frequenza è il risultato di un processo di classificazione: si individuano le h*k classi formate dalla coppia di modalità (x_i , y_j); si attribuisce ciascuna delle N unità statistiche alla classe corrispondente alla coppia di modalità osservata su quella unità; si contano le unità che sono state assegnate ad ogni classe.

Notazione → $(x_i , y_j , n_{ij})_{i=1,2,...,h;\ j=1,2,..,k}$ schematizzato in una tabella

	y_1	y_2	…	y_j	…	y_k	tot
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1k}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2k}	$n_{2\bullet}$
…	…	…	…	…	…	…	…
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{ik}	$n_{i\bullet}$
…	…	…	…	…	…	…	…
x_h	n_{h1}	n_{h2}	…	n_{hj}	…	n_{hk}	$n_{h\bullet}$
tot	$n_{\bullet 1}$	$n_{\bullet 1}$	…	$n_{\bullet j}$	…	$n_{\bullet k}$	N

dove $n_{i\bullet} = \sum_{j=1}^{k} n_{ij}$; $n_{\bullet j} = \sum_{i=1}^{h} n_{ij}$; $N = \sum_{i=1}^{h}\sum_{j=1}^{k} n_{ij} = \sum_{i=1}^{h} n_{i\bullet} = \sum_{j=1}^{k} n_{\bullet j}$

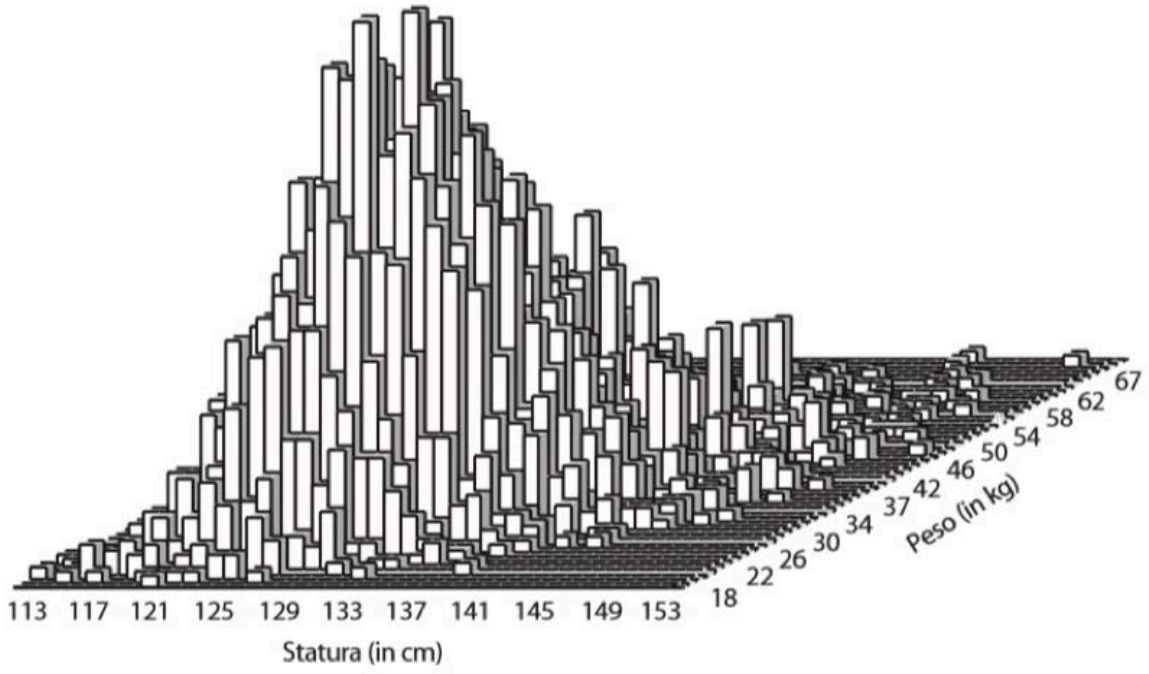
$(x_i , n_{i\bullet})_{i=1,2,...,h}$ → **distribuzione marginale di X**

$(y_j , n_{\bullet j})_{j=1,2,...,k}$ → **distribuzione marginale di Y**

$(x_i , n_{ij})_{i=1,2,...,h}$; $(y_j , n_{ij})_{j=1,2,...,k}$ → **distribuzioni condizionate**

Rappresentazione grafica

Individui classificati per statura (cm) e peso (kg)



Vedi costruzione in excel

Analisi dell'associazione tra due caratteri

Domanda: tra i due caratteri X e Y esiste una qualche forma di legame?

dipendenza logica → quando sono note a priori relazioni di causa/effetto di un carattere dall'altro.

> ***Nelle "scienze dure" le relazioni di causa/effetto sono espresse da "leggi" derivanti da teorie, magari scaturite dall'esame sperimentale dei dati, ma poi formalizzate in modelli generali.***
> ***Ad esempio, la differenza di potenziale elettrico ai due capi di un conduttore di resistenza data dipende dall'intensità di corrente che lo attraversa (legge d Ohm).***

indipendenza logica → quando nessuna teoria può giustificare la relazione di causa effetto.

Con l'analisi statistica non si pretende di individuare leggi che definiscono i legami tra variabili ma, più semplicemente, di verificare l'esistenza o meno di regolarità nell'associazione tra le modalità osservate dei due caratteri.

dipendenza (o interdipendenza) in senso statistico quando a certe modalità di un carattere tendono ad associarsi particolari modalità dell'altro e, quindi, la conoscenza della modalità di uno, diciamo di X, consente di fare migliori previsioni sulla modalità di Y che si troverà associata a quella.
E si parlerà di dipendenza di Y da X, quando si assume che X sia il carattere dominante: ad esempio, il consumo (Y) dipende dal reddito (X). Quando, invece, ai due caratteri si assegna lo stesso ruolo si parla di interdipendenza. Per contro

indipendenza in senso statistico quando la conoscenza delle modalità di un carattere non permette di migliorare la previsione sulla modalità dell'altro.

Indici di associazione

Gli aspetti da misurare possono essere due

- **Intensità del legame**
- **Direzione del legame (se le modalità sono almeno ordinabili**

Il tipo di indice utilizzabile dipende dalla scala con cui sono espresse le modalità dei due caratteri

nominale → connessione
ordinale → cograduazione
mista → dipendenza in media →
di intervallo o rapporto → correlazione

X nomin/ordinale
Y interv/rapporto

osservazione: le situazioni possono essere graduate in base al contenuto informativo delle scale di misura in cui sono espressi i due caratteri. La graduatoria (dal contenuto informativo meno ricco al più ricco) è la seguente

entrambi nominale → uno nominale e l'altro ordinale → entrambi ordinale → uno nominale/ordinale e l'altro di intervallo/rapporto → entrambi di intervallo/rapporto.

Gli indici costruiti per misurare l'associazione ad un livello della graduatoria possono sempre essere adottati per tutti i casi che seguono, mentre il viceversa non è possibile

DUE SCHEMI MENTALI DI RIFERIMENTO

(mutuati dall'analisi matematica)

- **Dipendenza funzionale unidirezionale (*bidirezionale*): ad ogni valore di X corrisponde un solo valore di Y *(e viceversa)* → y=f(x) *(f(.) monotona)***
- **Indipendenza funzionale: ad ogni valore di X corrisponde sempre lo stesso valore di Y → y=f(x)=k**

Indici di associazione bastati solo sulle frequenze

(qualunque tipo di carattere)

Data una distribuzione doppia di frequenza, possiamo immaginare una distribuzione interna delle frequenze che identifica una situazione di "legame perfetto"?

	y_1	y_2	Y_3	tot
x_1	40	-	-	40
x_2	-	-	30	30
X_3	-	30	-	30
tot	40	30	30	100

Ad ogni modalità di X corrisponde una sola modalità di Y e viceversa

Perfetta interdipendenza (dip. bilaterale) tra Y e X

	y_1	y_2	Y_3	tot
x_1	40	-	-	40
x_2	-	30	30	60
tot	40	30	30	100

Ad ogni modalità di Y corrisponde una sola modalità di X ma non vale il viceversa

Perfetta dipendenza di X da Y

	y_1	y_2	tot
x_1	40	-	40
x_2	-	30	30
X_3	-	30	30
tot	40	60	N

Ad ogni modalità di X corrisponde una sola modalità di Y ma non vale il viceversa

Perfetta dipendenza di Y da X

Le tre tabelle identificano una situazione che corrisponde al concetto di "dipendenza funzionale"

E l'indipendenza?

La situazione più vicina al concetto di indipendenza funzionale si ha quando al variare delle modalità di un carattere, diciamo X, non variano le distribuzioni condizionate di frequenza relativa di Y

	y_1	y_2	Y_3	tot
x_1	30	15	5	50
x_2	18	9	3	30
X_3	12	6	2	20
tot	60	30	10	100

	y_1	y_2	Y_3	tot
x_1	0,5	0,5	0,5	0,5
x_2	0.3	0.3	0.3	0.3
X_3	0.2	0.2	0.2	0.2
tot	1	1	1	1

Le distribuzioni condizionate di frequenza relativa di X non variano al variare di Y

	y_1	y_2	Y_3	tot
x_1	0,6	0,3	0,1	1
x_2	0,6	0,3	0,1	1
X_3	0,6	0,3	0,1	1
tot	0,6	0,3	0,1	1

Le distribuzioni condizionate di frequenza relativa di Y non variano al variare di X

In generale, si dice che due caratteri sono indipendenti in distribuzione quando le distribuzioni condizionate di frequenza relative sono uguali tra loro e quindi uguali a quelle marginali corrispondenti.

$$\frac{n_{ij}}{n_{\bullet j}} = \frac{n_{i\bullet}}{N} \;\ll==\gg\; \frac{n_{ij}}{n_{i\bullet}} = \frac{n_{\bullet j}}{N}$$

Da cui risulta che quando i due caratteri X e Y sono indipendenti, la frequenza assoluta della generica cella *" i-j "* (ovvero il numero di unità che presentano simultaneamente le modalità x_i e y_j) è pari a

$$n^*_{ij} = \frac{n_{i\bullet}\, n_{\bullet j}}{N}$$

Le frequenza così definite vengono dette frequenze teoriche di indipendenza (possono essere numeri non interi)

Se le frequenze effettive n_{ij} non coincidono con quelle teoriche, Si può affermare che tra i due caratteri X e Y sussiste "un certo grado di dipendenza".

Come misurarlo?

Si definisce "contingenza" La quantità

$$c_{ij} = (n_{ij} - n_{ij}^*)$$

che segnala quanto la frequenza della cella "ij" si discosta dalla situazione di indipendenza.

Per come sono costruite le contingenze, la loro somma è sempre pari a 0 (la somma delle frequenze teoriche e delle frequenze effettive è sempre N).

Considerando i quadrati delle contingenze, resi relativi rapportandoli alle frequenze toriche e sommandoli per tutte le h×k celle, si ottiene una misura della "distanza" della distribuzione effettiva da quella teorica di indipendenza (quindi, un indicatore diretto di dipendenza)

$$\chi^2 = \sum_{i=1}^{h} \sum_{j=1}^{k} \frac{c_{ij}^2}{n_{ij}^*}$$

Chi-quadrato di Pearson

dipende da N e dal numero (h×k) delle celle della tabella

da cui

$$\phi^2 = \frac{\chi^2}{N}$$

Contingenza quadratica media

Normalizzato rispetto a N dipende dal numero (h×k) delle celle della tabella

e infine

$$V = \sqrt{\frac{\phi^2}{min[h-1; k-1]}}$$

V di Cramér (0≤V≤1)

Dove min[h-1;k-1] è il massimo valore che ϕ^2 assume in caso di perfetta dipendenza

Osservazione: Valori degli indici maggiori di 0 segnalano una situazione di "non indipendenza" ma non spiegano il perché dell'associazione.

La giustificazione della presenza di una associazione tra i due caratteri è legata alle conoscenze che si hanno del fenomeno indagato e alle interpretazioni logiche che si possono dare.

Talvolta, tra due caratteri X e Y si riscontra una associazione che non si riesce a giustificare e che deriva dalla presenza di un carattere non osservato, che influenza sia X che Y. Si parla in questo caso di

ASSOCIAZIONE SPURIA

(ES: numero di auto/ora che attraversano il ponte di Brooklyn a varie ore della giornata e livello dell'acqua del Tamigi a Londra alle stesse ore)

Un esempio numerico: si intervistano 270 individui statunitensi chiedendo: "Se si potesse rivotare per le presidenziali, chi voterebbe?"

	black	White	tot
Trump	90	40	130
Clinton	60	25	85
non vote	40	15	55
	190	80	270

	black	white	tot
Trump	90	40	130
Clinton	60	25	85
non vote	40	15	55
tot	190	80	270

91,48=(130x190)/270

91,48	38,52
59,81	25,19
38,70	16,30

0.02=[(90-91,48)/91,48]^2

0,02	0,06
0,00	0,00
0,04	0,10

chi-quadrato	0,23
phi_quadrato	0,00085
V di Cramer	0,02915

L'indice chi-quadrato risulta pari a 0,23 e l'indice V è 0,02915, segnalando una associazione, sia pure debole, tra i due caratteri "razza" e "intenzioni di voto".

In realtà i dati provengono dall'unione delle due tabelle sotto riportate, in ciascuna delle quali i due caratteri "razza" e "intenzioni di voto" sono indipendenti.

È quindi il carattere "genere", non considerato nella prima tabella, il responsabile dell'associazione spuria prima rilevata.

maschi			
	black	white	tot
Trump	60	30	90
Clinton	30	15	45
non vote	10	5	15
	100	50	150

femmine			
	black	white	tot
Trump	30	10	40
Clinton	30	10	40
non vote	30	10	40
	90	30	120

Misura della dipendenza in media
(dipendenza di un carattere quantitativo da un altro di qualunque tipo)

Sia data la distribuzione doppia

$$(x_i , y_j , n_{ij})_{i=1,2,\ldots,h;\ j=1,2,..,k}$$

Supponiamo che il carattere Y sia quantitativo discreto (X qualunque).

$(y_j , n_{ij})_{j=1,2,\ldots,k}$ **è la distribuzione di Y condizionata alla modalità x_i del carattere X e la sua media e varianza possono essere indicate con**

$$\bar{y}_{X=x_i} = \frac{\sum_j y_j n_{ij}}{n_{i\bullet}} \quad ; \quad \sigma^2_{Y|X=x_i} = \frac{1}{n_{i\bullet}} \sum_j \left(y_j - \bar{y}_{X=x_i} \right)^2 n_{ij}$$

$(y_j , n_{\bullet j})_{j=1,2,\ldots,k}$ **è la distribuzione marginale di Y ed ha media e varianza**

$$\bar{y} = \frac{\sum_j y_j n_{\bullet j}}{N}$$

$$\sigma^2_Y = \frac{1}{N} \sum_j (y_j - \bar{y})^2 n_{\bullet j} = \frac{1}{N} \sum_i \sum_j (y_j - \bar{y})^2 n_{ij}$$

Se accade che le medie delle distribuzioni condizionate di Y sono tutte uguali tra loro e, quindi, uguali alla media della distribuzione marginale, si dice che Y è indipendente in media da X.

Come misurare la dipendenza in media?

La varianza di Y gode della proprietà di scomponibilità

$$\sigma_Y^2 = \frac{1}{N}\sum_i \sum_j \left[\left(y_j - \overline{y}_{X=x_i}\right) + \left(\overline{y}_{X=x_i} - \overline{y}\right)\right]^2 n_{ij}$$ *Var(Y) =*

$$\sigma_Y^2 = \frac{1}{N}\sum_i \sum_j \left(y_j - \overline{y}_{X=x_i}\right)^2 n_{ij}$$ *A +*

$$+ \frac{1}{N}\sum_i \sum_j \left(\overline{y}_{X=x_i} - \overline{y}\right)^2 n_{ij}$$ *B +*

$$+ 2\,\frac{1}{N}\sum_i \sum_j \left(y_j - \overline{y}_{X=x_i}\right)\left(\overline{y}_{X=x_i} - \overline{y}\right) n_{ij}$$ *C*

$$A = \frac{1}{N}\sum_i n_{i\bullet}\,\sigma^2_{Y|X=x_i} = \textit{varianza interna gruppi}$$

$$B = \frac{1}{N}\sum_i \left(\overline{y}_{X=x_i} - \overline{y}\right)^2 n_{i\bullet} = \textit{varianza tra gruppi}$$

$$C = 2\frac{1}{N}\sum_i \left(\overline{y}_{X=x_i} - \overline{y}\right)\sum_j \left(y_j - \overline{y}_{X=x_i}\right) n_{ij} = 0$$

Si può misurare la dipendenza di Y da X con il rapporto di correlazione:

$$\eta^2_{y|x} = \frac{var\ tra}{var\ tot} = \frac{\frac{1}{N}\sum_i \left(\overline{y}_{X=x_i} - \overline{y}\right)^2 n_{i\bullet}}{\frac{1}{N}\sum_j \left(y_j - \overline{y}\right)^2 n_{\bullet j}}$$

L’indice varia tra

0 → varianza tra gruppi=0
medie delle distribuzioni di Y condizionate dalle modalità di X tutte uguali tra loro e, quindi, uguali alla media generale di Y

1 → varianza interna ai gruppi=0
in ogni gruppo tutte le unità hanno lo stesso valore di Y (nessuna variabilità)

ed esprime quanta parte della variabilità di Y è da attribuire al legame con il carattere X (tutta la variabilità sarebbe dovuta ad Y se X non presentasse variabilità)

Esempio numerico: resa per ettaro rilevata su 10 appezzamenti di terreno sottoposti a differenti tipi di concimazione

	X = concime			
Y = Resa per ha	A	B	C	totale
23	-	-	1	1
24	-	1	2	3
25	-	2	1	3
26	1	1	-	2
28	1	-	-	1
totale	2	4	4	10

$$\bar{y} = \frac{23 \times 1 + 24 \times 3 + 25 \times 3 + 26 \times 2 + 2\,8 \times 1}{10} = 25$$

$$\sigma_y^2 = \frac{1}{10} \times [(23-25)^2 \times 1 + (24-25)^2 \times 3 + (25-25)^2 \times 3$$

$$+(26-25)^2 \times 2 + (28-25)^2 \times 1] = \frac{18}{10} = 1{,}8$$

$$\bar{y}_{x=A} = \frac{[26\times1+28\times1]}{2} = 27 \qquad \sigma^2_{y|x=A} = \frac{[(26-27)^2\times1+(28-27)^2\times1]}{2} = 1$$

$$\bar{y}_{x=B} = 25 \qquad \sigma^2_{y|x=B} = 0,5$$

$$\bar{y}_{x=C} = 24 \qquad \sigma^2_{y|x=C} = 0,5$$

$$Var\ Tra = \frac{(27-25)^2\times2+(25-25)^2\times4+(24-25)^2\times4}{10} = \frac{12}{10} = 1{,}2$$

$$Var\ Entro = \frac{(1\times2)+(0{,}5\times4)+(0{,}5\times4)}{10} = \frac{6}{10} = 0{,}6$$

Rapporto di correlazione = 1,2/1,8 = 0,67

Il 67% della variabilità riscontrata nelle rese per ettaro è imputabile alla diversa concimazione

Confronto tra i due concetti di indipendenza

Che relazione esiste tra indice V di Cramer e rapporto di correlazione (ossia tra misura della dipendenza in distribuzione e della dipendenza in media)?

Il concetto di indipendenza in distribuzione è più stringente: se due caratteri sono indipendenti in distribuzione lo sono anche in media ma non vale il viceversa.

vedi foglio excel
"indip in media VS indip in distribuz"

Misura della covarianza/correlazione
(interdipendenza tra caratteri quantitativi)

Sia data la distribuzione doppia

$$(x_i , y_j , n_{ij})_{i=1,2,...,h;\ j=1,2,..,k}$$

riferita a due caratteri quantitativi, che supponiamo discreti *(se di tipo continuo si assume che le unità in ogni classe presentino tutte modalità pari al valore centrale della classe stessa)*

Siano $\overline{x}$ e $\overline{y}$ le medie e si consideri il prodotto

	$(x_i - \overline{x})$	$(y_i - \overline{y})$	
$x_i \leq \overline{x}\ e\ y_i \leq \overline{y}$	–	–	= +
$x_i \geq \overline{x}\ e\ y_i \geq \overline{y}$	+	+	= +
$x_i \leq \overline{x}\ e\ y_i \geq \overline{y}$	–	+	= –
$x_i \leq \overline{x}\ e\ y_i \leq \overline{y}$	+	–	= –

La media di tutti i prodotti è detta covarianza

$$cov(x,y) = \frac{\sum_i \sum_j (x_i - \overline{x})(y_i - \overline{y}) n_{ij}}{N}$$

e segnala la concordanza o discordanza riscontrata

concordanza → a modalità sopra media di X tendono a corrispondere modalità sopra media di Y

vedi foglio excel "Covarianza"