TABELLE A DOPPIA ENTRATA

SI CHIAMA STATISTICA UNIVARIATA QUELLA CHE SI OCCUPA DELL'ANALISI DEI DATI PROVENIENTI DALLA RILEVAZIONE DI UN SOLO CARATTERE IN UNA DATA POPOLAZIONE. SI CHIAMA INVECE STATISTICA BIVARIATA QUELLA CHE RILEVA DUE CARATTERI X ED Y CONTEMPORANEAMENTE, ALLO SCOPO DI METTERE IN LUCE EVENTUALI CORRELAZIONI TRA X ED Y.
RILEVIAMO DUNQUE I DATI RELATIVI A DUE CARATTERI X ED Y SU CIASCUNA DELLE N UNITÀ CHE COMPONGONO LA POPOLAZIONE: IL RISULTATO È UNA TABELLA DI DATI GREZZI, COME QUELLA A SINISTRA.

UNITÀ STATISTICHE	MODALITÀ DI X	MODALITÀ DI Y
1	x_1	y_1
2	x_2	y_2
...	...	...
N	x_N	y_N

NELLA STATISTICA UNIVARIATA, SI COSTRUISCE UNA TABELLA CON LA DISTRIBUZIONE DI FREQUENZE; IN QUELLA BIVARIATA, SI COSTRUISCE INVECE UNA TABELLA A DOPPIA ENTRATA, CHE RIPORTA LE FREQUENZE CON CUI SI MANIFESTANO LE COPPIE DI MODALITÀ OSSERVATE. COSTRUIAMO CIOÈ UNA TABELLA DI $(K+1)$ RIGHE E $(H+1)$ COLONNE, RIPORTANDO NELLA PRIMA RIGA LE MODALITÀ $y_1, y_2, \dots y_N$ DI Y E NELLA PRIMA COLONNA LE MODALITÀ $x_1, x_2, \dots x_N$ DI X. NELLA CASELLA ALL'INCROCIO TRA LA i-ESIMA RIGA E LA j-ESIMA COLONNA RIPORTIAMO LA FREQUENZA ASSOLUTA DELLA COPPIA $(x_i; y_j)$, CHE DI SOLITO VIENE INDICATA CON LA SCRITTURA $f(x_i; y_j)$. LE FREQUENZE ASSOLUTE DI TALI COPPIE SONO DETTE FREQUENZE CONGIUNTE, E LA TABELLA COSÌ COSTRUITA VIENE DETTA DISTRIBUZIONE DOPPIA DI FREQUENZE. LA TABELLA A DOPPIA ENTRATA COSÌ COSTRUITA SI COMPLETA CON UN'ULTIMA RIGA, IN CUI VENGONO RIPORTATE LE SOMME DELLE FREQUENZE DI CIASCUNA COLONNA, E UN'ULTIMA COLONNA, DOVE VENGONO RIPORTATE LE SOMME DELLE FREQUENZE DI CIASCUNA RIGA. QUESTE RIGHE E COLONNE CONTENGONO LE COSIDDETTE DISTRIBUZIONI MARGINALI, CIOÈ LE DISTRIBUZIONI CHE SI AVREBBERO SE CIASCUNO DEI DUE CARATTERI FOSSE PRESO SINGOLARMENTE.
AD ESEMPIO, RILEVIAMO L'ETÀ E IL MESTIERE DI UN GRUPPO DI PERSONE, OTTENENDO LA SEGUENTE TABELLA DI DATI GREZZI E LA SEGUENTE TABELLA A DOPPIA ENTRATA:

NOME	ETÀ	LAVORO
ADRIANO	18	STUDENTE
BENEDETTA	20	OPERAIO
CAMILLO	19	STUDENTE
DARIO	20	SEGRETARIO
ELENA	18	OPERAIO

X \ Y	OPERAIO	SEGRETARIO	STUDENTE	TOT
18	1	0	1	2
19	1	0	0	1
20	1	1	0	2
TOT	3	1	1	5

LE FREQUENZE MARGINALI SI INDICANO DI SOLITO CON $f(x_1), f(x_2), \dots, f(x_N)$, $f(y_1), f(y_2), \dots, f(y_N)$. ANCHE IN QUESTO CASO È POSSIBILE OTTENERE LE FREQUENZE MARGINALI RELATIVE DI X E Y, COSTRUENDO UNA TABELLA CON I RAPPORTI TRA LE SINGOLE FREQUENZE MARGINALI E IL NUMERO TOTALE DI UNITÀ DELLA NOSTRA POPOLAZIONE.

(→)

Facciamo ancora riferimento alla tabella presa come esempio. Se consideriamo la prima modalità di X, $x_1 = 18$, leggiamo come si distribuisce il carattere Y (il lavoro) tra le unità della popolazione che manifestano la modalità x_1 di X. Per questo si dice che tale riga rappresenta la <u>distribuzione condizionata</u> di Y rispetto alla modalità x_1 di X. Insomma, fissare l'attenzione su una singola riga o colonna della tabella escludendo le altre, significa restringersi a una "sottopopolazione" che presenta una data modalità di X o di Y. Ciascuna di queste righe o colonne, presa singolarmente, rappresenta perciò una particolare distribuzione condizionata. È anche possibile costruire le distribuzioni condizionate relative, dividendo le frequenze congiunte che appartengono alla distribuzione condizionata considerata con i corrispondenti totali di riga o di colonna, come si vede nell'esempio qui a sinistra.

X	$X\|y_1$	$X\|y_1$ rel.
18	1	1/3
19	1	1/3
20	1	1/3
TOT	3	1

$X|y_1$ è il simbolo per indicare le frequenze di X condizionate dalla modalità y_1 di Y.

Dati due caratteri X ed Y, per stabilire se l'una dipende o meno dall'altra l'idea è quella di confrontare le distribuzioni di X condizionate alle modalità di Y con la distribuzione marginale di X (che esprime il comportamento di X considerato singolarmente). Se c'è indipendenza, c'è da aspettarsi che il condizionamento di X alle modalità di Y non abbia alcun effetto, cioè che le distribuzioni condizionate si mantengano uguali a quella marginale. Attenzione, però: le frequenze marginali si riferiscono all'<u>intera</u> popolazione, mentre le frequenze condizionate si riferiscono soltanto alla sottopopolazione che presenta la modalità rispetto cui stiamo condizionando. Non sarebbe perciò corretto eseguire il confronto tra due frequenze <u>assolute</u>: il confronto deve essere eseguito tra frequenze <u>relative</u>.

Il carattere X si dice <u>indipendente</u> dal carattere Y se le distribuzioni condizionate relative di X rispetto alle modalità di Y sono uguali alla distribuzione marginale relativa di X.

Si dimostra che due caratteri X e Y, dei quali sono state osservate le modalità $x_1, x_2, \dots, x_N$ e $y_1, y_2, \dots, y_N$ su una popolazione costituita da N unità sono indipendenti tra loro se e solo se:

$$f(x_i; y_j) = \frac{f(x_i) \cdot f(y_j)}{N} \qquad \begin{pmatrix} i = 1, 2, \dots, N \\ j = 1, 2, \dots, N \end{pmatrix}$$

In pratica, X e Y sono indipendenti tra loro se ogni frequenza congiunta è uguale al prodotto delle corrispondenti frequenze marginali, divise per N.

(→)

(⟶)

LA FORMULA RIPORTATA È SIMMETRICA; CIOÈ, SE X È INDIPENDENTE DA Y, Y È INDIPENDENTE DA X. LE FREQUENZE CONGIUNTE CHE VERIFICANO LA CONDIZIONE DI INDIPENDENZA STATISTICA SI DICONO <u>FREQUENZE TEORICHE DI INDIPENDENZA</u>, PER DISTINGUERLE DA QUELLE EFFETTIVAMENTE OSSERVATE, E LE SI INDICA CON $f'(x_i; y_j)$. AD OGNI TABELLA CHE RAPPRESENTA UNA DOPPIA DISTRIBUZIONE DI FREQUENZE OSSERVATE, È ALLORA POSSIBILE ASSOCIARE UNA TABELLA TEORICA DI INDIPENDENZA, COSTRUITA TENENDO FISSE LE DISTRIBUZIONI MARGINALI, E SOSTITUENDO LE FREQUENZE CONGIUNTE OSSERVATE CON QUELLE TEORICHE DI INDIPENDENZA. L'INDIPENDENZA TRA I DUE CARATTERI IN ESAME È VERIFICATA SE E SOLO SE LA TABELLA TEORICA DI INDIPENDENZA COINCIDE CON LA TABELLA DELLE FREQUENZE OSSERVATE.

CONSIDERIAMO AD ESEMPIO LA TABELLA QUI SOTTO A SINISTRA, CHE SI RIFERISCE A UNA POPOLAZIONE DI 100 UNITÀ, E OSSERVIAMO (A DESTRA) LA TABELLA TEORICA DI INDIPENDENZA.

X \ Y	y_1	y_2	TOT
x_1	35	20	55
x_2	8	17	25
x_3	14	6	20
TOT	57	43	100

X \ Y	y_1	y_2	TOT
x_1	$\frac{57 \cdot 55}{100} = 31$	$\frac{43 \cdot 55}{100} = 24$	55
x_2	$\frac{57 \cdot 25}{100} = 14$	$\frac{43 \cdot 25}{100} = 11$	25
x_3	$\frac{57 \cdot 20}{100} = 12$	$\frac{43 \cdot 20}{100} = 8$	20
TOT	57	43	100

POICHÉ LE DUE TABELLE NON COINCIDONO, VUOL DIRE CHE I DUE CARATTERI X ED Y <u>NON SONO</u> STATISTICAMENTE INDIPENDENTI.

È IMPORTANTE OSSERVARE CHE, MENTRE LE FREQUENZE CONGIUNTE OSSERVATE SONO ESPRESSE DA <u>NUMERI INTERI</u>, LE FREQUENZE TEORICHE DI INDIPENDENZA IN GENERALE SONO ESPRESSE DA <u>NUMERI DECIMALI</u>, RISULTANDO DA RAPPORTI TRA I PRODOTTI DELLE FREQUENZE MARGINALI E IL NUMERO COMPLESSIVO DI UNITÀ DELLA POPOLAZIONE. LA SITUAZIONE DI PERFETTA INDIPENDENZA STATISTICA SI PUÒ PERCIÒ REALIZZARE SOLO NEL CASO IN CUI TUTTE LE FREQUENZE TEORICHE DI INDIPENDENZA SONO ESPRESSE DA NUMERI INTERI. QUESTO CI DICE CHE È MOLTO, MOLTO DIFFICILE CHE LA TABELLA DELLE FREQUENZE OSSERVATE COINCIDA ESATTAMENTE CON LA TABELLA TEORICA DI INDIPENDENZA. QUESTA SITUAZIONE VA QUINDI CONSIDERATA COME <u>IDEALE</u>, ED È PIUTTOSTO IMPORTANTE CAPIRE DI QUANTO I DATI REALI SI DISCOSTANO DA ESSA. PERCIÒ OCCORRE STABILIRE UNA MISURA QUANTITATIVA DEL GRADO DI INDIPENDENZA DI X E DI Y. AD ESEMPIO SI PUÒ USARE IL <u>TEST DEL CHI QUADRO</u>.