Lezione 7
Analisi delle relazioni tra due caratteri

POLO UNIVERSITARIO
FONDAZIONE
PORTOGRUARO CAMPUS

PROF. ROBERTO COSTA

SCIENZE DELL'EDUCAZIONE - STATISTICA SOCIALE (305SF)

Dove eravamo rimasti

Nella scorsa lezione abbiamo iniziato a parlare di relazioni tra due variabili e di analisi statistica bivariata.

Abbiamo visto come possiamo costruire una tabella di contingenza (o a doppia entrata), che mette assieme l'andamento di due variabili.

Ci sono dei dubbi?

Le relazioni tra due caratteri

Prendiamo in considerazione due variabili X e Y.

Da un punto di vista logico possiamo conoscere che tra queste due variabili esista una relazione di causa ed effetto. Ovvero che al variare di una corrisponda una variazione, in un preciso verso, dell'altra.

Questa relazione può derivare da leggi (ad esempio nella fisica), ovvero da conoscenze acquisite.

Nell'ambito della statistica sociale la relazione è frutto di ragionamenti logico-deduttivi.

Secondo Marradi, con l'analisi bivariata si cerca di individuare **forma**, **forza** e **direzione** delle relazioni tra due variabili.

Forma

Supponiamo di avere 2 diverse distribuzioni di frequenza congiunte in due regioni italiane.

Titolo studio/ Reddito	Basso	Medio	Alto	Totale
Basso	180	10	10	200
Medio	10	180	10	200
Alto	10	10	180	200
Totale	200	200	200	600

Titolo studio/ Reddito	Basso	Medio	Alto	Totale
Basso	130	50	20	200
Medio	60	80	60	200
Alto	10	70	120	200
Totale	200	200	200	600

Nel primo caso vi è un'evidente relazione tra il reddito e il titolo di studio, mentre nel secondo chi ha un titolo di studio medio si distribuisce in modo più omogeneo nei vari livelli di reddito.

Queste sono le **forme** della relazione. Nel primo caso è evidente il **segno**, ovvero l'andamento.

Forma

Per parlare di segno è necessario che le variabili siano, perlomeno, qualitative ordinate.

Se al crescere di una variabile cresce anche l'altra si parla di **relazione positiva**.

Al contrario si parla di **relazione negativa**.

Forza

Le due distribuzioni viste prima, presentano la stessa forma, ma non la stessa forza.

La forza sarebbe massima se, tornando al nostro esempio, tutti gli individui con titolo di studio basso avessero un reddito basso, quelli con titolo di studio medio, un reddito medio e quelli con titolo di studio alto un reddito alto.

Titolo di studio/reddito	Basso	Medio	Alto	Totale
Basso	200	0	0	200
Medio	0	200	0	200
Alto	0	0	200	200
Totale	200	200	200	600

Titolo di studio/reddito	Basso	Medio	Alto	Totale
Basso	90	60	50	200
Medio	60	80	60	200
Alto	50	60	90	200
Totale	200	200	200	600

Le relazioni tra due caratteri

Qualora si riscontri la presenza di una direzione di una relazione, nelle scienze sociali si vuole capire anche se esista un nesso di causalità tra le variabili e, in caso affermativo, quale variabile influisce sull'altra (**direzione causale**).

Nell'ambito delle scienze sociali l'individuazione di una direzione causale non è così semplice per svariati motivi:

- Le tecniche di analisi statistica non permettono di stabilire la direzione causale, ma solo di accertarla,
- Le relazioni fra variabili hanno un carattere «tendenziale»,
- L'individuazione di una direzione causale è resa difficile dal fatto che molte relazioni sono bidirezionali, ovvero le variabili si influenzano reciprocamente,
- Se individuiamo una direzione causale, non è detto sia facile ricostruire il meccanismo causale
- Qualsiasi variabile dipendente, può dipendere da molti elementi.

Le relazioni tra due caratteri

Nell'ambito delle scienze sociali quando studiamo un legame tra due variabili, in realtà misuriamo solo una parte di questa relazione, lasciando un'altra parte non spiegata.

Ad esempio, possiamo supporre che tra il reddito (X) e la spesa per cultura e spettacoli (Y) vi sia una dipendenza logica, con Y dipendente da X.

Noi però coglieremo statisticamente solo una parte di questa relazione, poiché Y potrà dipendere anche da altre variabili (come ad es. l'età, il livello di istruzione, ecc.).

Non vale la relazione inversa (il reddito non dipende dalla spesa per cultura e spettacoli).

Il rapporto di causa ed effetto risulta unidirezionale e asimmetrico, in statistica viene studiato attraverso **l'analisi della dipendenza**.

Variabili dipendenti e indipendenti

L’attribuzione di una direzione causale implica l’individuazione, tra le variabili messe in relazione di una dipendente e una indipendente.

La variabile indipendente influisce sulla variabile dipendente senza esserne, a sua volta, influenzata.

Ad esempio:

Titolo di studio e utilizzo di internet

Persone di 6 anni e più per utilizzo di Internet e frequenza di utilizzo. Anno 2022, valori %

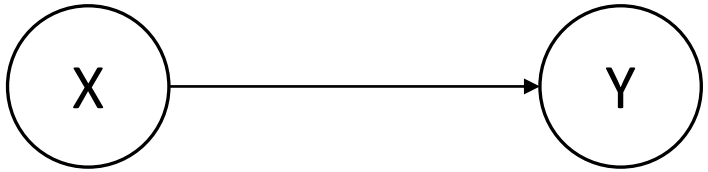
	sì	no
licenza di scuola elementare, nessun titolo di studio	49,8	50,2
licenza di scuola media	76,3	23,7
diploma	89,7	10,3
laurea e post-laurea	94,9	5,1
totale	**78,5**	**21,5**

(Fonte: Istat, Multiscopo sulle famiglie: aspetti della vita quotidiana)

Rappresentazioni

Solitamente si rappresenta con X la variabile indipendente e Y la variabile dipendente.

La struttura tipo di una relazione bivariata si può quindi rappresentare così:



Dove le variabili sono rappresentate da cerchi e la direzione causale da una freccia.

Tipi di variabile e tecniche di analisi

Come abbiamo già visto, la scelta delle tecniche da adottare dipende dal tipo di variabili analizzate (qualitative, quantitative, ecc.).

La scelta delle tecniche di analisi bivariata e multivariata segue il seguente schema:

Tipo di variabile dipendente	Tipo di variabile indipendente	
	Qualitativa	Quantitativa
Qualitativa	Tabulazione incrociata Regressione logistica	Regressione logistica
Quantitativa	Regressione semplice Regressione multipla	Regressione semplice Regressione multipla

Tratto da: Statistica per la ricerca sociale, Corbetta P., Gasperoni G. e Pisati M. (Il Mulino, 2001)

Il concetto di correlazione

Due variabili si dicono **correlate** quando i valori di una variabile Y tendono a seguire quelli di un'altra variabile X con una certa regolarità.

Abbiamo già fatto diverse ipotesi di variabili correlate: reddito e spesa per cultura e spettacoli, titolo di studio e utilizzo di internet.

La correlazione però non è necessariamente solo positiva.

Il concetto di correlazione

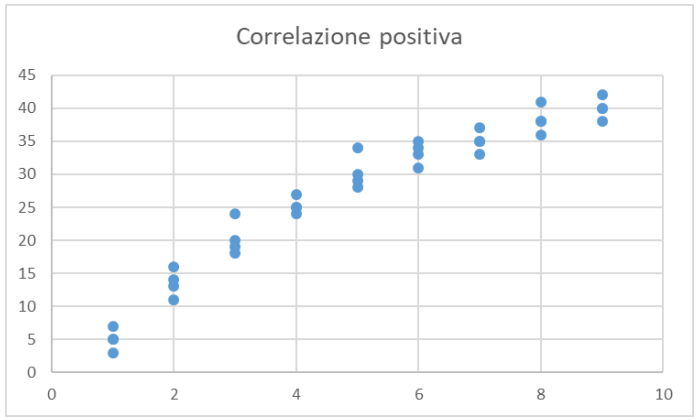
Ci possiamo trovare in diverse situazioni:

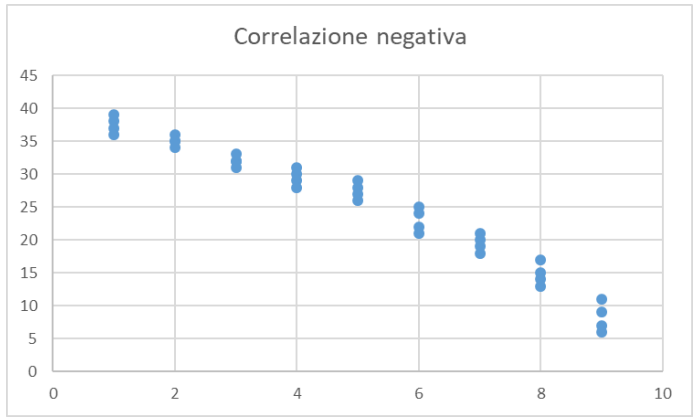
Correlazione positiva (o concordanza): due variabili crescono contemporaneamente (ad esempio reddito e spesa per cultura e spettacoli).

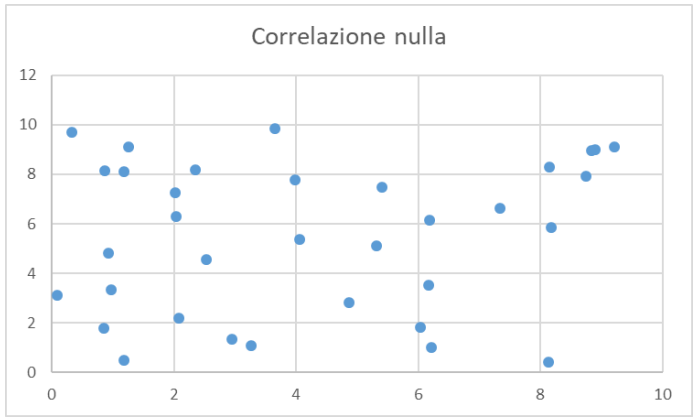
Correlazione negativa (o discordanza): una variabile cresce mentre l'altra diminuisce (ad esempio, titolo di studio e tasso di disoccupazione).

Correlazione nulla (o incorrelazione): due variabili non hanno la tendenza a crescere o diminuire contemporaneamente.

Il concetto di correlazione





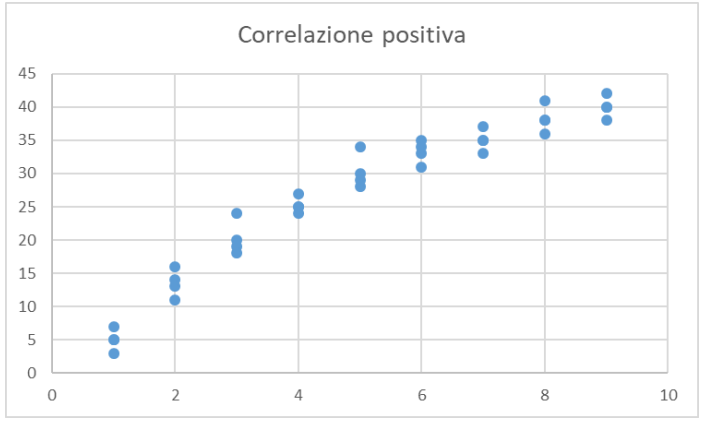


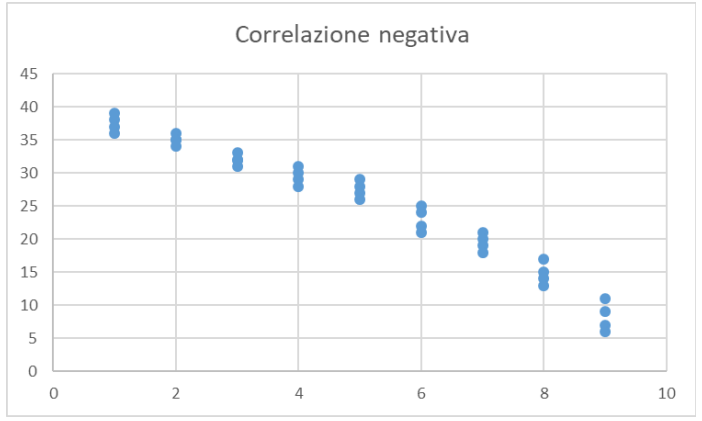
Il concetto di correlazione

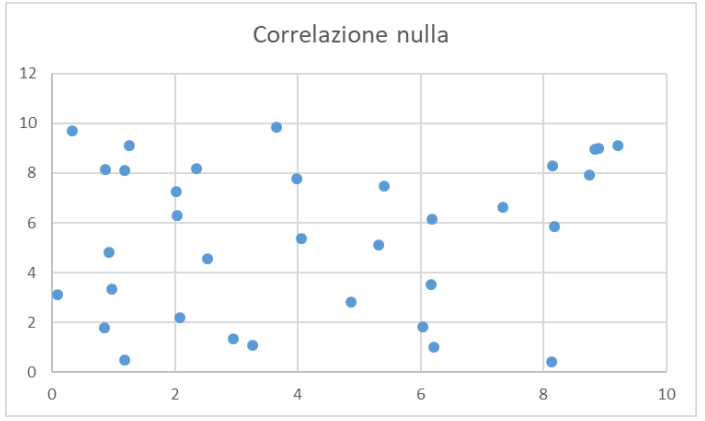
Correlazione positiva
I dati tendono a distribuirsi lungo una linea retta crescente, dove X e Y crescono contemporaneamente.

Correlazione negativa
I dati tendono a distribuirsi lungo una linea retta decrescente, dove mentre X cresce Y diminuisce.

Correlazione nulla
I dati tendono a distribuirsi in maniera casuale.







Calcolo della covarianza

Consideriamo due variabili quantitative X e Y.

Possiamo calcolare gli scarti S rispetto alle medie di X e di Y.

$S(x_i) = x_i - M(x)$

$S(y_j) = y_j - M(y)$

Possiamo quindi calcolare il prodotto degli scarti:

$S(x_i)\, S(y_j) = [x_i - M(x)]\, [y_j - M(y)]$

Calcolo della covarianza

La covarianza è la media del prodotto degli scarti $S(x_i)$ e $S(y_i)$.

Partiamo dagli scarti:

$S(x_i) = x_i - M(x)$

$S(y_j) = y_j - M(y)$

Possiamo calcolare la covarianza:

$\text{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n} S(xi)\, S(yj) = \frac{1}{n}\sum_{i=1}^{n} [xi - M(x)]\, [yj - M(y)]$

L’indice di correlazione di Pearson

Al fine di misurare in modo più chiaro la correlazione tra due variabili possiamo utilizzare l’indice di correlazione di Pearson (detto anche **coefficiente di correlazione lineare**).

È un indice che esprime un'eventuale relazione di linearità tra le variabili.

L’indice di correlazione di Pearson, può assumere un valore compreso +1 e -1, dove +1 corrisponde alla perfetta correlazione lineare positiva, 0 corrisponde a un'assenza di correlazione lineare e -1 corrisponde alla perfetta correlazione lineare negativa.

L’indice di correlazione di Pearson

Date due variabili statistiche X e Y, **l'indice di correlazione di Pearson** è definito come la loro covarianza divisa per il prodotto delle deviazioni standard delle due variabili.

$$\rho = \frac{Cov(XY)}{\sigma_x \, \sigma_y}$$

Se ρ > 0 si ha correlazione positiva tra X e Y

Se ρ = 0 non si ha correlazione tra X e Y

Se ρ < 0 si ha correlazione negativa tra X e Y

Il coefficiente di correlazione di Pearson

Per quanto riguarda la «forza» della relazione:

Se $0 < |\rho_{xy}| < 0,3$ si ha correlazione debole tra X e Y

Se $0,3 < |\rho_{xy}| < 0,7$ si ha correlazione moderata tra X e Y

Se $|\rho_{xy}| > 0,7$ si ha correlazione forte tra X e Y

Esercitiamoci (1)

Chiedo a un campione di 8 bambini quanti libri hanno letto per svago nell'ultimo anno e il voto di italiano che hanno conseguito in pagella alla fine del quadrimestre.

Queste sono le risposte ottenute:

Bambino	1	2	3	4	5	6	7	8
Libri letti	8	12	10	15	2	9	7	5
Voto italiano	7	9	8	10	6	8	7	6

Esercitiamoci (1)

Bambino	1	2	3	4	5	6	7	8
Libri letti	8	12	10	15	2	9	7	5
Voto italiano	7	9	8	10	6	8	7	6

Rispondiamo ad alcune domande:

Che tipo di variabili sono «Libri letti per svago» e «Voto in italiano»?

Per queste due variabili posso calcolare media e mediana?

Esercitiamoci (1)

Bambino	1	2	3	4	5	6	7	8
Libri letti	8	12	10	15	2	9	7	5
Voto italiano	7	9	8	10	6	8	7	6

Calcoliamo per ogni variabile (analisi univariata):

Gli indici di tendenza centrale: media aritmetica e mediana.

Gli indici di variabilità: range (campo di variazione), varianza e scarto quadratico medio, coefficiente di variazione.

Per le due variabili congiuntamente (analisi bivariata):

Il coefficiente di correlazione di Pearson.

Esercitiamoci (1)

Inseriamo i dati in un foglio di calcolo:

Bambino	Libri letti	Voto in italiano	x-M(x)	y-M(y)	(x-M(x))*(y-M(y))
1	8	7			
2	12	9			
3	10	8			
4	15	10			
5	2	6			
6	9	8			
7	7	7			
8	5	6			

Esercitiamoci (1)

Quali formule mi servono per risolvere i quesiti di analisi univariata:

Media aritmetica	$M(x) = \frac{\sum_{i=1}^{k} n_i}{N}$
Mediana (numero pari di casi)	Me = (N/2 + (N/2 + 1)) /2
Range	$x_{max} - x_{min}$
Varianza	$\sigma^2 = \Sigma(x_i - m)^2/N$
Scarto quadratico medio	$\sigma = \frac{\sqrt{\Sigma(x_i - m)^2}}{N}$
Coefficiente di variazione	$C_v = \frac{\sigma}{m_x}$

Esercitiamoci (1)

Quali formule mi servono per risolvere i quesiti di analisi bivariata:

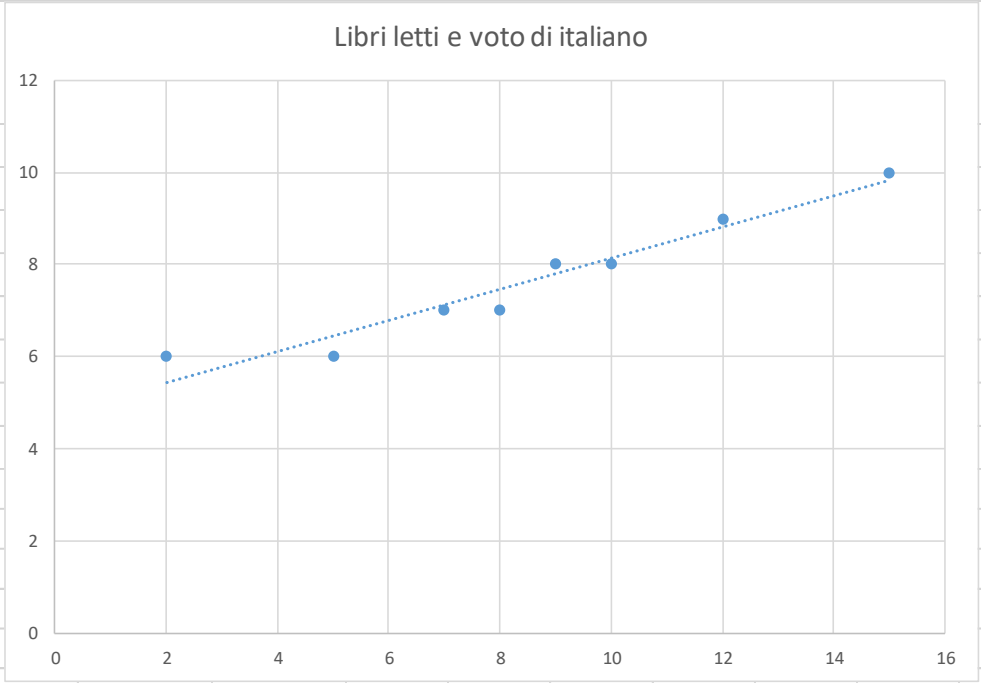
Coefficiente di correlazione di Pearson $\rho = \frac{Cov(XY)}{\sigma_x \, \sigma_y}$

Come calcolo la covarianza:

Cov (X,Y) = $\frac{1}{n}\sum_{i=1}^{n} S(x_i)\, S(y_j) = \frac{1}{n}\sum_{i=1}^{n} [x_i - M(x)]\, [y_j - M(y)]$

Esercitiamoci (1)

Bambino	Libri letti	Voto italiano	(x - m(x))	(y - m(y))	(x - m(x)) * (y - m(y))
1	8	7	-0,5	-0,625	0,3125
2	12	9	3,5	1,375	4,8125
3	10	8	1,5	0,375	0,5625
4	15	10	6,5	2,375	15,4375
5	2	6	-6,5	-1,625	10,5625
6	9	8	0,5	0,375	0,1875
7	7	7	-1,5	-0,625	0,9375
8	5	6	-3,5	-1,625	5,6875
					38,5
Media	8,50	7,63			
Varianza	14,25	1,73			
Deviazione standard	3,77	1,32			
Covarianza	4,81				
Indice di correlazione	0,97				



Le relazioni tra due caratteri

Prendiamo in considerazione due variabili X e Y.

Come posso procedere nel caso di variabili qualitative, oppure in presenza di una tabella di contingenza (non dei microdati)?

Partiamo da un esempio: supponiamo di aver chiesto agli studenti di tre scuole medie quale percorso pensano di intraprendere dopo la licenza media.

Scuola di orig./scuola scelta	Liceo	Istituto tecnico	Totale
Scuola A	12	48	60
Scuola B	16	64	80
Scuola C	8	32	40
Totale	36	144	180

Indipendenza in distribuzione

Dobbiamo partire dalle frequenze condizionate.

Scuola di orig./scuola scelta	Liceo	Istituto tecnico	Totale
Scuola A	0,33	0,33	0,33
Scuola B	0,44	0,44	0,44
Scuola C	0,22	0,22	0,22
Totale	1,00	1,00	1,00

Distr. condiz. di X (scuola di orig.) rispetto a Y (scuola scelta) X|Y

Scuola di orig./scuola scelta	Liceo	Istituto tecnico	Totale
Scuola A	0,20	0,80	1,00
Scuola B	0,20	0,80	1,00
Scuola C	0,20	0,80	1,00
Totale	0,20	0,80	1,00

Distr. condiz. di Y (scuola scelta) rispetto a X (scuola di orig.) Y|Y

Se vogliamo capire se due caratteri sono indipendenti devo analizzare le frequenze relative condizionate di X rispetto a Y e di Y rispetto a X.

Indipendenza in distribuzione

Nel nostro esempio possiamo vedere come le distribuzioni relative condizionate di X rispetto alle modalità di Y sono tutte uguali tra di loro e rispetto alla frequenza relativa marginale di X.

Anche le distribuzioni relative condizionate di Y rispetto alle modalità di X sono uguali tra di loro e uguali alla frequenza relativa marginale di Y.

Ci troviamo nel caso di **indipendenza statistica in distribuzione**.

Scuola di orig./scuola scelta	Liceo	Istituto tecnico	Totale
Scuola A	0,33	0,33	0,33
Scuola B	0,44	0,44	0,44
Scuola C	0,22	0,22	0,22
Totale	1,00	1,00	1,00

Scuola di orig./scuola scelta	Liceo	Istituto tecnico	Totale
Scuola A	0,20	0,80	1,00
Scuola B	0,20	0,80	1,00
Scuola C	0,20	0,80	1,00
Totale	0,20	0,80	1,00

Indipendenza statistica in distribuzione

Formalizziamo il ragionamento che abbiamo fatto prima.

X è statisticamente indipendente da Y se le *h* distribuzioni di frequenza relativa di X condizionate alle modalità di Y sono uguali alla frequenza relativa marginale di X:

$\frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{N}$ per ogni i = 1, …, k e ogni j = 1, …, h

L'indipendenza è simmetrica, quindi

$\frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{N}$ per ogni i = 1, …, k e ogni j = 1, …, h

Possiamo parlare di indipendenza tra X e Y senza dover specificare una direzione.

Indipendenza statistica in distribuzione

In sintesi, X e Y sono indipendenti se le distribuzioni di frequenza relativa marginale di X|Y sono uguali alla distribuzione di frequenza relativa marginale di X e se le distribuzioni marginali di Y|X sono uguali alla distribuzione di frequenza relativa marginale di Y.

Partendo dalla definizione di indipendenza, dire che X e Y sono statisticamente indipendenti significa affermare che:

$$n_{ij}^{*}=\frac{n_{i.}n_{.j}}{N}$$

Ovvero ogni frequenza assoluta congiunta è pari al prodotto del totale della riga i per il totale della colonna j diviso per il numero totale di unità N.

La frequenza assoluta congiunta che otteniamo si chiama **frequenza teorica**, che otterremmo nel caso di indipendenza assoluta tra i caratteri X e Y.

n_{ij}^{*}= frequenza teorica = $\frac{n_{i.}n_{.j}}{N}$

Dipendenza perfetta in distribuzione

Vediamo ora un caso opposto di **dipendenza perfetta** in distribuzione di due variabili X e Y.

Partiamo da un esempio: supponiamo di aver chiesto agli studenti di una scuola da quale rione provengono.

Un carattere Y dipende perfettamente da X quando a ogni modalità di X è associata una sola modalità di Y, ovvero quando per ogni riga i c'è una sola colonna j dove $n_{ij} \neq 0$.

Scuola/rione	Rione 1	Rione 2	Totale
Scuola A	150		150
Scuola B		200	200
Scuola C	100		100
Totale	250	200	450

Interdipendenza perfetta in distribuzione

Si parla di **interdipendenza perfetta** tra due caratteri X e Y quando a ogni modalità di X è associata una sola modalità di Y e, allo stesso tempo, a ogni modalità di Y è associata una sola modalità di X.

Questo significa che per ogni riga i c'è solo una colonna j dove $n_{ij} \neq 0$ e viceversa.

Possiamo trovare l'interdipendenza perfetta solo nel caso di una tabella quadrata, ovvero con lo stesso numero di righe e colonne.

Scuola/rione	Rione 1	Rione 2	Rione 3	Totale
Scuola A	150			150
Scuola B		200		200
Scuola C			100	100
Totale	150	200	100	450

Dipendenza e indipendenza in distribuzione

Pe riassumere, abbiamo visto tre diversi casi che possiamo verificare in una tabella a doppia entrata:

1. Indipendenza in distribuzione $\frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{N}$ per ogni I = 1, …, k e ogni j = 1, …, h
2. Dipendenza perfetta in distribuzione
3. Interdipendenza perfetta in distribuzione

Esercitiamoci

Abbiamo chiesto a 12 alunni di una scuola elementare il numero di ore alla settimana dedicate al gioco all'aperto:

6 2 9 13 13 14 0 6 13 2 4 1

- Che tipo di variabile è?

- Calcoliamo alcuni valori centrali:

Media

Moda

Mediana

Esercitiamoci

Abbiamo chiesto a 12 alunni di una scuola elementare il numero di ore alla settimana dedicate al gioco all'aperto:

6 2 9 13 13 14 0 6 13 2 4 1

- Calcoliamo il campo di variazione

- Calcoliamo la varianza

- Calcoliamo lo scarto quadratico medio