SESTA UNITA'

Il concetto di correlazione

Fino a questo momento ci siamo interessati alle varie statistiche che ci consentono di descrivere la distribuzione dei punteggi di una data variabile e di collegare le varie statistiche introdotte alla interpretazione dei singoli punteggi della distribuzione. Tuttavia molti problemi di indagine vanno al di là della semplice descrizione di una sola variabile nei suoi vari aspetti. Siamo, in effetti, spesso chiamati a determinare la relazione esistente tra due o più variabili. Per esempio, una scuola è interessata a conoscere la relazione esistente tra i punteggi di ammissione in una data classe ed il successivo rendimento medio degli studenti di quella classe. Ancora, i genitori con più alto quoziente di intelligenza hanno figli ugualmente dotati? Esiste relazione tra classe socio-economica e recidività nel crimine?

Associazioni tra caratteri

Nello studio delle relazioni che possono essere prese in considerazione occorre in primo luogo distinguere tra caratteri qualitativi e caratteri quantitativi. Ricordiamo come le articolazioni dei caratteri qualitativi (scale nominali e scale ordinali) sono denominate modalità del carattere, mentre le variazioni dei caratteri quantitativi danno luogo a variabili (continue o discrete a secondo del valore numerico che possono assumere: numeri reali o numeri interi)

Per evidenziare le relazioni esistenti tra due caratteri qualitativi si usano normalmente tabelle a doppia entrata. Se denominiamo *A* e *B* i due caratteri, le *i* modalità della *A* vengono indicate con a_i, mentre le *j* modalità della *B* si indicano con b_j. Le frequenze delle osservazioni che rilevano l'associazione tra due caratteri vengono in genere indicate con n_{ij}. Tabelle di questo tipo sono anche denominate **tabelle di contingenza** (dal latino *contingo*) e le relazioni esistenti tra i caratteri **associazioni dei caratteri**.

Ecco un esempio di tabella generica nella quale il carattere A è articolato secondo 3 modalità e il carattere B è articolato secondo 2 modalità.

	Modalità b_1	**Modalità b_2**	
Modalità a_1	n_{11}	n_{12}	Σ_{1j}
Modalità a_2	n_{21}	n_{22}	Σ_{2j}
Modalità a_3	n_{31}	n_{32}	Σ_{3j}
	Σ_{i1}	Σ_{i2}	Σ_{ij}

Le osservazioni di frequenza n_{11} si riferiscono al numero delle unità prese in considerazione per la modalità a_1 associata alla modalità b_1, e così via. Vedremo in seguito come si esaminano le eventuali relazioni esistenti tra i due caratteri.

Esempio: studenti che appartengono alla FSE o alla FT e che frequentano i corsi che nell'anno accademico 2004-2005 sono insegnati dallo stesso docente (dati fittizi).

	Facoltà FSE	**Facoltà FT_2**	**Totali**
Pedagogia Gen	n_{11}= 134	n_{12} = 12	Σ_{1j} = 146
Statistica	n_{21}= 92	n_{22} = 1	Σ_{2j} = 93
Didattica Gen_3	n_{31}= 21	n_{32} = 5	Σ_{3j} = 26
Totali	Σ_{i1}= 247	Σ_{i2} = 18	Σ_{ij} = 265 = N

Questa tabella di contingenza può essere facilmente trasformata indicando nella varie caselle le proporzioni numeriche corrispondenti. A esempio nel corso di pedagogia generale (prima riga) gli studenti della Facoltà FSE rappresentano il 92% del totale. Questo valore si trova in questo modo 134/146 x 100 = 0,92. La tabella risultante è la seguente.

	Facoltà FSE	**Facoltà FT_2**	**Totali parziali**
Pedagogia Gener.	51%	5%	56%
Statistica	34%	0% (0,4)	34%
Didattica Gener.	8%	2%	10%
Totali parziali	93%	7%	100%

Nel caso in cui i caratteri siano due e di tipo quantitativo è possibile utilizzare la consueta forma di rappresentazione grafica della relazione esistente tra le variabili prese in considerazione. Se queste vengono denominate *X* e *Y* si può costruire un diagramma di tipo cartesiano con ascissa *X* e ordinata *Y*. Ne deriva quello che è stato denominato in inglese *scatterplot*: un diagramma nel quale sono segnati i punti individuati dai valori dell'ascissa e dell'ordinata.

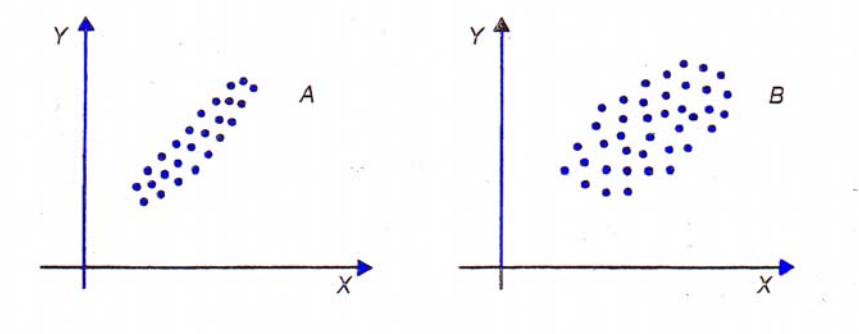

Fig.1 - *Esempi di diagrammi di associazione tra due caratteri quantitativi*

Nel caso *A* si può notare una certa tendenza delle modalità del carattere *X* a essere associate in maniera significativa a quelle dell'altro carattere *Y*. Nel caso *B* la situazione è più incerta.

Naturalmente è possibile anche rappresentare le associazioni tra caratteri qualitativi e quantitativi. Ecco un esempio nel quale il carattere qualitativo *A* è articolato secondo le modalità "regioni italiane", mentre il carattere quantitativo *B* è costituito dalla numerosità della popolazione residente articolata per classi di età.

Regioni	Classe di età-anni			
	25-44	45-64	65 e più	Totale
Piemonte	1.259.443	1.178.266	801.592	4.297.989
Valle d'Aosta	37.020	31.163	20.428	118.456
Lombardia	2.742.931	2.389.143	1.407.575	8.910.451
Trentino Alto Adige	283.000	214.348	142.423	908.667
Bolzano-Bozen	*140.205*	*103.534*	*63.645*	*449.055*
Trento	*142.795*	*110.814*	*78.778*	*459.612*
Veneto	1.371.953	1.117.341	727.160	4.422.290
Friuli Venezia Giulia	346.623	322.546	240.504	1.191.248
Liguria	458.616	465.725	379.829	1.663.696
Emilia Romagna	1.143.081	1.062.566	817.910	3.922.604
Toscana	1.001.218	948.767	726.748	3.526.031
Umbria	229.300	215.397	169.375	822.480
Marche	411.182	367.086	285.477	1.441.031
Lazio	1.579.595	1.321.699	798.677	5.193.233
Abruzzo	364.353	301.344	227.349	1.267.694
Molise	93.803	76.812	61.612	332.155
Campania	1.690.220	1.204.606	693.573	5.745.761
Puglia	1.178.731	898.927	545.193	4.075.802
Basilicata	175.877	135.279	95.393	610.699
Calabria	598.263	441.501	300.176	2.076.128
Sicilia	1.450.305	1.114.842	736.577	5.082.697
Sardegna	511.005	374.298	223.501	1.659.466
Italia	16.926.519	14.181.656	9.401.072	57.268.578

Tab. 1 - *Popolazione residente in Italia al 1 gennaio1995 distribuita per regione e per classe di età*

La correlazione tra due variabili

Al fine di esprimere in maniera quantitativa l'intensità del legame tra due variabili è necessario infatti calcolare un **indice di correlazione**. Esistono vari tipi di indici di correlazione; la scelta di un indice in particolare dipende in generale da vari fattori: 1) il tipo di livello di misurazione di ciascuna delle due variabili; 2) la natura della distribuzione sottostante (continua o discreta); 3) le caratteristiche della distribuzione dei punteggi nel diagramma (lineare o non lineare).

Vengono qui presentati due indici di correlazione: la ***r* di Pearson**, altrimenti detta **coefficiente di correlazione**, generalmente utilizzata per variabili misurate con scale di intervallo o di rapporto e **la *r* di Spearman**, cioè il **coefficiente di correlazione per le graduatorie**, utilizzato nel caso di dati disposti in successioni ordinate.

Indipendentemente dal tipo di indice di correlazione che si intende usare, tutti gli indici presentano alcune caratteristiche comuni.

1. I due insiemi di punteggi sono associati agli stessi individui od eventi, o a soggetti diversi ma associati tra loro da uno specifico punto di vista.

2. I valori dei vari indici di correlazione variano tra -1 e +1; ambedue i valori estremi rappresentano relazioni perfette tra le variabili, mentre 0 rappresenta l'assenza di relazione. Questo almeno finché consideriamo relazioni di tipo lineare.

3. Una relazione positiva significa che gli individui che ottengono valori elevati in una variabile tendono ed ottenere valori elevati sulla seconda variabile. Ed è vero anche

viceversa, cioè coloro che hanno bassi valori su una variabile tendono ed avere bassi valori sulla seconda variabile.

4. Una relazione negativa sta a indicare che a bassi punteggi su una variabile corrispondono alti punteggi sull'altra variabile.

Per esplorare visivamente la relazione esistente tra due variabili abbiamo visto che è utile rappresentarla mediante opportuni **diagrammi di correlazione** (**scatterplot**). Sull'asse orizzontale, o asse delle ascisse, vengono riportati i valori della prima variabile, la variabile *X*, mentre sull'asse verticale, o asse delle ordinate, vengono riportati i valori della seconda variabile, la variabile *Y*. Consideriamo una indagine condotta negli anni ottanta da Snyder e Simpson (1984). Essi ipotizzarono che la fedeltà in una relazione sentimentale dipendesse da una caratteristica di personalità descritta come tendenza a dipendere nel proprio comportamento dal contesto o dalle circostanze incontrate. Applicando opportuni strumenti di rilevazione essi ottennero una serie di dati. I valori della prima variabile, la *X*, potevano andare da 0 a 25. La seconda variabile, la *Y*, teneva conto del numero di mesi di frequentazione costante dello stesso partner. I risultati sono riassunti nella tabella seguente.

Soggetto	X	X^2	Y	Y^2
1	14	196	8	64
2	19	361	4	16
3	12	144	4	16
4	9	81	14	196
5	15	225	8	64
6	6	36	14	196
7	21	441	10	100
8	14	196	14	196
9	12	144	14	196
10	15	225	4	16
11	14	196	10	100
12	11	121	10	100
13	7	49	16	256
14	6	36	18	324
15	15	225	6	36
16	15	225	8	64
17	10	100	10	100
18	16	256	8	64
19	16	256	14	196
20	10	100	12	144
$N = 20$	$\Sigma X = 257$	$\Sigma X^2 = 3613$	$\Sigma Y = 206$	$\Sigma Y^2 = 2444$
	$\overline{X} = 12.85$		$\overline{Y} = 10.30$	
	$s_X = 3.94$		$s_Y = 4.01$	

X		Y
	0–4	444
9766	5–9	68888
44422100	10–14	0000244444
9665555	15–19	68
1	20–21	

Tab. 6.2 - *Distribuzione dei dati delle variabili X e Y e relativa rappresentazione mediante ramo e foglie.*

Nella rappresentazione grafica della relazione esistente tra le due serie di dati si individua un insieme di punti. Ogni punto rappresenta i valori di *X* e *Y* di ogni soggetto preso in considerazione. In altre parole, ogni punto è individuato da due valori: il punteggio individuale nella variabile *X* e il punteggio individuale, della stessa persona, nella variabile *Y*.

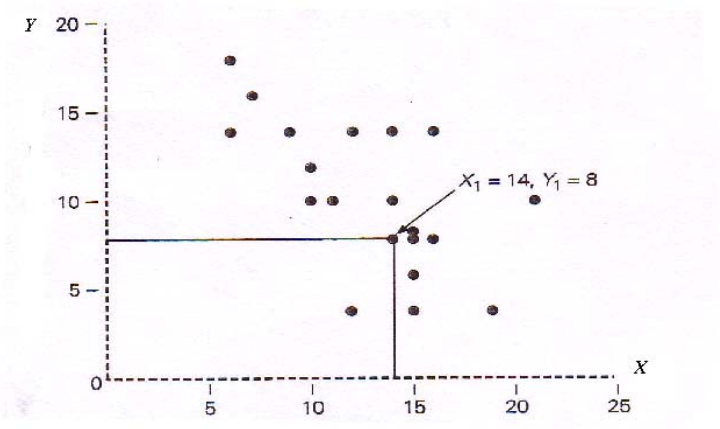

Fig. 6.2 - *Diagramma di dispersione della correlazione esistente tra le variabili X e Y.*

Si nota facilmente che la correlazione è negativa, nel senso che a un alto valore nella dipendenza dal contesto o dalle circostanze corrisponde un valore basso nel numero di mesi e, viceversa, a un valore basso nella *X* corrisponde un valore alto nella *Y*.

Ecco alcuni possibili diagrammi di correlazione e relativo valore del coefficiente *r*.

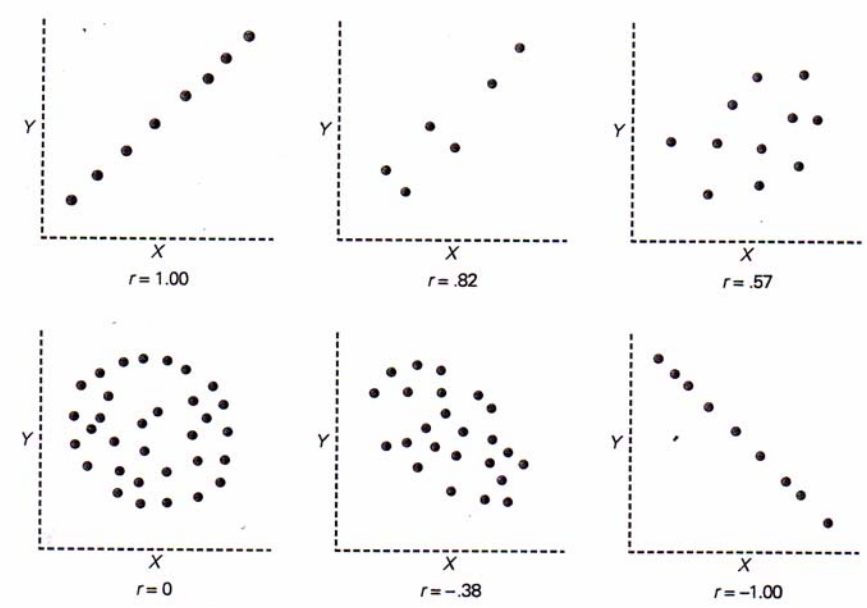

Fig. 6.3 - *Alcuni possibili diagrammi e relativo coefficiente di correlazione.*

Il coefficiente *r* e i punteggi z

Un alto valore positivo del coefficiente di correlazione *r* di Pearson indica che ciascun individuo dovrebbe ottenere più o meno lo stesso punteggio *z* su ambedue le variabili. Nel caso di correlazione positiva perfetta (*r* = 1,00), ogni individuo otterrà esattamente lo stesso punteggio *z* su ambedue le variabili. Analogamente, nel caso di una *r* altamente negativa, ciascun individuo dovrebbe ottenere approssimativamente lo stesso punteggio *z* sulle due variabili, ma di segno opposto.

Ricordando che i punteggi *z* rappresentano una misura della posizione relativa di un punteggio individuale in una data distribuzione, cioè una *z* altamente positiva rappresenta un punteggio elevato relativamente a tutta la distribuzione, mentre una *z* altamente negativa rappresenta un basso punteggio relativamente a tutta la distribuzione, possiamo generalizzare il significato del coefficiente *r*.

> *Il coefficiente di correlazione r di Pearson*[1]*, rappresenta il grado di concordanza o discordanza della posizione dello stesso individuo in due variabili.*

La modalità attraverso la quale si calcolano i punteggi *z* ci assicura che le unità di misura utilizzate per trasformare le variabili *X* e *Y* in punteggi *z* non debbono essere necessariamente uguali nel calcolo di *r* . L'indipendenza di *r* dall'unità di misura e dall'ampiezza del campo di variazione ci permette di analizzare la relazione tra variabili qualsiasi. Potremmo al limite calcolare la correlazione tra la lunghezza dell'alluce e il quoziente di intelligenza nel caso ne fossimo interessati.

Si noti altresì, come abbiamo già sottolineato, che i punteggi *z* di ogni soggetto in ciascuna variabile sono gli stessi nel caso di correlazione positiva massima. Nel caso in cui si rovesci l'ordine di successione dei punteggi in ciascuna variabile, i punteggi *z* saranno ancora gli stessi, sebbene di segno opposto. In quest'ultimo caso la correlazione sarà una correlazione negativa perfetta (*r* = -1,00).

Se moltiplichiamo i valori associati dei punteggi *z* e quindi li sommiamo, otteniamo il massimo solo quando il coefficiente di correlazione è 1,00. Infatti, mano a mano che il coefficiente di correlazione si avvicina allo zero, la somma dei prodotti delle variabili *z* associate si avvicina anche essa a zero. Da notare che quando la correlazione è perfetta, la somma dei prodotti dei valori *z* appaiati è uguale ad n, cioè al numero della coppie stesse. Queste constatazioni ci portano ad una delle formule esplicite per il calcolo di *r*, tra le tante tra loro algebricamente equivalenti:

$$r = \frac{\sum \left(z_x z_y \right)}{N}$$

La formula ora presentata è di difficile elaborazione in quanto il calcolo effettivo di *r* implica la valutazione dei punteggi *z* per ogni soggetto. Potete immaginare il lavoro da fare non appena si superi un n pari a 50, cosa piuttosto frequente nelle scienze sociali.

[1] Spesso il coefficiente di Pearson è denominato coefficiente di Bravais-Pearson. Soprattutto nel contesto francese è poi chiamato semplicemente coefficiente di Bravais.

Per questi motivi, si danno diverse altre formule di calcolo della *r*. Qui verranno presentate due formule per il calcolo di *r:* la formula dello scarto medio e la formula dei punteggi originali.

Calcolo del coefficiente di correlazione *r* di Pearson

Metodo dello scarto medio

Il metodo dello scarto medio per il calcolo del coefficiente dl correlazione *r* di Pearson, come la precedente formula basata sui punteggi *z*, non è di uso comune presso i ricercatori nella scienze del comportamento, perché implica un maggior tempo nonché una maggior mole di calcoli rispetto ad altre formule. La si presenta ugualmente in quanto permette di chiarire meglio le proprietà del coefficiente di correlazione *r*. Tuttavia quando *N* è piccolo si mostra ancora conveniente, a meno che naturalmente non si disponga di un computer. La formula è

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \cdot \sum y^2}}$$

Metodo dei punteggi originali

Utilizzando i punteggi originali nella formula per il calcolo delle somme dei quadrati si ha:

$$\sum x^2 = \sum X^2 - \frac{\left(\sum X\right)^2}{N}$$

$$\sum y^2 = \sum Y^2 - \frac{\left(\sum Y\right)^2}{N}$$

Per analogia, otteniamo che la somma dei prodotti assodati è esprimibile in termini di punteggi originali:

$$\sum xy = \sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{N}$$

Nel calcolo della *r* di Pearson, basato sui punteggi originali, si ha la possibilità di scegliere tra il calcolo delle precedenti quantità prese separatamente e quindi introdotte dalla formula, ed il calcolo di *r* effettuato direttamente sui dati originali, come nelle formule:

$$r = \frac{\frac{\sum XY}{N} - \bar{X}\bar{Y}}{\sqrt{\left(\frac{\sum X^2}{N} - \bar{X}^2\right)\left(\frac{\sum Y^2}{N} - \bar{Y}^2\right)}}$$

Si può notare che il denominatore della formula consiste nel prodotto tra gli scostamenti quadratici medi delle variabile *X* e della variabile *Y* cioè $s_x s_y$. Ne ricaviamo una nuova formula per il calcolo di *r*, cioè:

$$r = \frac{\frac{\sum XY}{N} - \overline{X}\,\overline{Y}}{s_x s_y}.$$

Il numeratore della formula precedente è denominato *covarianza* ed è un indice del grado con cui le due variabili condividono una comune varianza. Essa è uguale a 1 quando vi è una perfetta correlazione; è uguale a 0 quando non ve n'è alcuna.

Le procedure per il calcolo di *r* direttamente dai dati originali sono riportate nella tabella 6.3. Come nel caso del metodo dello scarto medio, quasi tutte le altre procedure sono ormai familiari dalle precedenti formule per il calcolo dello scarto quadratico medio sui dati originali. La quantità ΣXY è ottenuta molto semplicemente moltiplicando ciascun valore delle *X* per il corrispondente valore della *Y* e quindi sommando i prodotti.

SOGGETTO	X	X^2	Y	Y^2	XY
A	1	1	7	49	7
B	3	9	4	16	12
C	5	25	13	169	65
D	7	49	16	256	112
E	9	81	10	100	90
F	11	121	22	484	242
G	13	169	19	361	247
	$\Sigma X = 49$	$\Sigma X^2 = 455$	$\Sigma Y = 91$	$\Sigma Y^2 = 1435$	$\Sigma XY = 775$
	$\overline{X} = 7$		$\overline{Y} = 13$		

$$s_x = \sqrt{\frac{455}{7} - (7)^2} = \sqrt{65 - 49} = \sqrt{16} = 4$$

$$s_y = \sqrt{\frac{1435}{7} - (13)^2} = \sqrt{205 - 169} = \sqrt{36} = 6$$

$$r = \frac{\frac{\Sigma XY}{n} - \overline{X}\,\overline{Y}}{s_x s_y} = \frac{\frac{775}{7} - (7)(13)}{(4)(6)}$$

$$= \frac{110{,}71 - 91}{24} = \frac{19{,}71}{24} = 0{,}82$$

Tab. 6.3 - *Procedura di calcolo del coefficiente di correlazione r di Pearson, basata sui dati originali*

Occorre precisare che quando si hanno coefficienti di correlazione vicini allo zero si è tentati fortemente di concludere che non esiste alcuna relazione tra e due variabili oggetto di studio. Tuttavia bisogna ricordare che il coefficiente di correlazione *r* di Pearson misura l'intensità della relazione lineare tra due variabili. Il fatto che non si trovi alcuna evidenza della esistenza di una qualche relazione tra le due variabili può essere dovuto e due fattori: 1) le variabili sono in effetti indipendenti completamente; 2) le variabili sono legate da una relazione non lineare. In quest'ultimo caso la *r* di Pearson non è più una misura valida del grado di correlazione tre le due variabili. Così, se riportiamo in un grafico i dati relativi

all'età ed alla capacità di attenzione in un campione di individui, dovremmo ottenere un grafico dal tipo della Fig. 6.4.

Fig. 6.4 - *Diagramma relativo alla correlazione non lineare di due variabili.*

In genere è possibile determinare l'eventuale allontanamento del dati dalla condizione di linearità semplicemente dando un'occhiata al grafico. Un allontanamento di piccole entità della linearità non influenzerà in generale la grandezza del coefficiente di correlazione. D'altra parte, là dove esiste una evidente curvilinearità, come nella Figura 6.3, un coefficiente di correlazione curvilinea rifletterà molto meglio l'intensità della relazione tra le due variabili in questione. In queste dispense non ci occuperemo di coefficienti di correlazione non lineare, tuttavia è bene che si sia coscienti di questa possibilità e pertanto, come prassi, si dovrebbe costruire un grafico prima comunque di calcolare le *r* di Pearson.

Punteggi ordinali e coefficiente *r* di Spearman

Supponiamo che voi siate insegnanti di scuola secondaria superiore e che per molti anni abbiate effettuato delle rilevazioni nelle vostre classi in base al sospetto che l'intelligenza sia strettamente legata alla personalità (nel senso di influenza sui compagni di classe). Nel tentativo di controllare questa ipotesi, disporrete delle stima del QI di tutti i ragazzi della classe. Tuttavia scoprite di non disporre di strumenti di misura della attitudine alla leadership nei vostri studenti e quindi di non avere modo di quantificare questo carattere. Ciononostante, mediante numerose osservazioni del comportamento dei ragazzi in situazioni di leadership, pensate di saper costruire una graduatoria dei vostri ragazzi secondo questa capacità dal più basso al più alto. I dati che ne risultano costituiscono naturalmente una scala ordinale. Nei casi in cui almeno una delle variabili sia una scala ordinale, si può utilizzare il coefficiente di correlazione r_{rho} di Spearman, o coefficiente di correlazione tra graduatorie. Il coefficiente di Spearman si rivela utile anche quando una sola delle due variabili costituisca una graduatoria, mentre l'altra può essere sia un ordinamento lineare che una scala ad intervalli o di rapporto. Tuttavia, prima di applicare il coefficiente è necessario che ambedue le misurazioni siano riportate in termini di graduatoria.
Coscienti che la conoscenza dell'intelligenza dei vostri ragazzi può infirmare la vostra stima della loro attitudine al comando, chiederete ad un vostro collega di ordinare egli stesso gli studenti secondo la loro attitudine alla leadership. Dopodiché ordinate gli studenti dal primo all'ultimo, secondo il loro quoziente di intelligenza, ottenendone una graduatoria. Il coefficiente di correlazione fra graduatorie richiede il calcolo delle differenze tra i posti nella graduatoria, il loro quadrato e quindi le somma dei quadrati. La formula che ne risulta è la seguente dove *D* = posto *X*-esimo - posto *Y*-esimo.

$$r_{rho} = 1 - \frac{6 \sum D^2}{n(n^2 - 1)}$$

Come dato di fatto, Σ*D* dovrebbe essere calcolato indipendentemente della sua utilità o meno per successivi calcoli, in quanto costituisce un utile controllo dell'accuratezza dei vostri calcoli fino a quel punto, essendo sempre Σ*D* uguale a zero. Se ottenete un valore diverso da zero dovete ricontrollare la vostra graduatoria nonché le successive differenze.

Sintesi

Si è presentato il concetto di correlazione e il calcolo di due coefficienti in particolare: la *r* di Pearson, utilizzato nel caso si disponga di scale intervallo o di rapporto, ed il coefficiente r_{rho}, generalmente usato nel caso di variabili ordinali.

Abbiamo visto che la correlazione ha a che fare con la misura della intensità del legame associativo tra due variabili, cioè di come esse tendono a variare simultaneamente.

L'espressione quantitativa del legame è data in termini di grandezza del coefficiente di correlazione. Esso può variare tra -1,00 e +1,00, dove gli estremi rappresentano legami associativi perfetti. Nel caso che il coefficiente sia zero, si ha una prova della assenza di relazione tra le due variabili. Abbiamo notato che le *r* di Pearson ha validità unicamente nel caso che le variabili siano legate linearmente. Nel caso di dati organizzati serialmente, cioè ordinali, il coefficiente di correlazione tra graduatorie di Spearman rappresenta l'esatta controparte di *r*. Sono state presentate altresì varie formule per il calcolo della *r* di Pearson.

Termini da ricordare

Associazione – Relazione esistente tra caratteri (in particolare di tipo qualitativo).

Correlazione - Relazione tra due variabili di tipo quantitativo.

Coefficiente di correlazione - Indice che esprime l'intensità del legame associativo tra due variabili.

Relazione negativa - Due variabili sono dette associate negativamente quando ad un alto punteggio nell'una corrisponde un basso punteggio nell'altra, e viceversa ad un basso punteggio nell'una corrisponde un alto punteggio nell'altra.

***r* di Pearson** - Coefficiente di correlazione per variabili misurate con scale di intervallo o di rapporto.

Covarianza – Indice del grado con cui due variabili condividono una comune varianza.

Relazione positiva - Due variabili sono associate positivamente quando ad un alto punteggio nell'una corrisponde un alto punteggio nell'altra, e viceversa, ad un basso punteggio nell'una corrisponde un basso punteggio nell'altra.

Diagramma - Strumento grafico atto a rappresentare la variazione di due variabili.

***r* di Spearman** - Coefficiente di correlazione utilizzato nel caso di dati in graduatoria.

Esercizi

1. La tabella mostra i punteggi ottenuti da un gruppo di 20 studenti agli esami di ammissione in un college (*X*) e un test di comprensione verbale.

Studente	Esami di ammissione *X*	Test di comprensione verbale *Y*
A	52	49
B	28	34
C	70	45
D	51	49
E	49	40
F	65	50
G	49	37
H	49	49
I	63	52
J	32	32
K	64	53
L	43	41
M	35	28
N	66	50
O	26	17
P	44	41
Q	49	29
R	28	17
S	30	15
T	60	55

a) Disegna un diagramma di correlazione dei seguenti dati e descrivi la relazione tra le due prove.
b) Calcola il coefficiente di correlazione tra queste due variabili. Quale coefficiente di correlazione sostiene la tua descrizione del diagramma di correlazione?

2. I dati seguenti sono relativi ai punteggi ottenuti da 10 studenti in un esame di statistica e la media dei voti riportati durante i quattro anni. Prepara il diagramma e calcola la *r* di Pearson.

Studente	Esame di Statistica *X*	Media finale alla Laurea *Y*
A	27	28
B	30	25
C	18	24
D	18	22
E	25	24
F	28	26
G	30	25
H	18	26
I	22	20
J	30	30

3. Spiega con parole tue il significato della correlazione.

4. In un gruppo di 50 individui abbiamo che $\Sigma z_x z_y$ è pari a 41,3. Qual è la correlazione tra le due variabili?