3. La matrice dei dati e le analisi preliminari

3.1 Introduzione

Per realizzare una analisi statistica concernente fenomeni aziendali, o di qualsiasi altra natura, non basta raccogliere dati, bisogna anche organizzarli in modo appropriato. Sia che i dati provengano da fonti secondarie (si veda il Capitolo 1) o da rilevazioni ad hoc (Capitolo 2), essi vanno organizzati in modo da rendere possibili le analisi statistiche, che spesso coinvolgono contemporaneamente una pluralità di variabili, come si vedrà più avanti (Capitoli 4 e 5).
Organizzare i dati in modo appropriato significa sistemarli in una matrice di dati, ovvero in una tavola composta da un certo numero di righe e di colonne. Nelle righe vengono in genere collocati gli oggetti, cioè le unità di osservazione, o unità statistiche, che possono essere individui, imprese, o anche modalità di un carattere, ad esempio la regione di residenza o il settore di attività delle imprese, ecc. Nelle colonne vengono invece collocati gli attributi delle unità statistiche, cioè le diverse variabili misurate su di esse.
Ad esempio, se si deve realizzare una analisi statistica di dati raccolti tramite una indagine campionaria sulle imprese, nelle righe della matrice dei dati verranno collocate le diverse imprese del campione, mentre nelle colonne verranno collocate le variabili che su ognuna di esse sono state rilevate.
Queste ultime possono essere di tipologie eterogenee per livello di misurazione: alcune quantitative, come il volume dei ricavi o il numero dei dipendenti; altre qualitative ordinali, come il titolo di studio del titolare dell'impresa; altre ancora qualitative sconnesse, come la forma giuridica. Le analisi che potranno essere compiute, gli indici statistici che potranno essere calcolati, naturalmente saranno diversi a seconda del tipo di variabili coinvolte.
In questo capitolo vengono illustrate le principali analisi preliminari che possono essere condotte su matrici di dati del tipo accennato. In particolare, verranno richiamate le analisi che possono essere condotte sui cosiddetti "profili di colonna" e quelle sui cosiddetti "profili di riga" della matrice. I profili di colonna si riferiscono alle distribuzioni delle singole variabili tra le unità statistiche, mentre i profili di riga descrivono le singole unità statistiche sulla base delle molteplici variabili su di esse rilevate.
Per quanto riguarda le analisi sui profili di colonna verranno tuttavia tralasciate le analisi cosiddette univariate, concernenti cioè le singole variabili, che sono oggetto di approfondita trattazione nei corsi di statistica di base. Verranno invece richiamate le principali analisi bivariate, concernenti in particolare il grado di associazione tra coppie di variabili presenti nella matrice dei dati. Con riferimento ai profili riga verranno invece richiamate le principali misure di distanza e similarità tra unità statistiche.
Gli indici di associazione o di distanza che verranno illustrati in questo capitolo costituiscono spesso il punto di partenza per le più complesse analisi statistiche che saranno oggetto dei prossimi due capitoli: per la formulazione dei modelli di regressione multivariata (Capitolo 4); per la realizzazione delle diverse analisi multidimensionali (Capitolo 5).
Una volta introdotta la matrice dei dati nella sua forma più comune (Paragrafo 3.2), il capitolo si apre con un richiamo ai principali problemi di qualità dei dati contenuti nella matrice, in particolare quelli derivanti dalla presenza di valori anomali (*outliers*) e dalla presenza di mancate risposte parziali (Paragrafo 3.3). Successivamente vengono illustrate le principali misure di associazione tra variabili, nei diversi casi di variabili qualitative, quantitative o miste (Paragrafo 3.4). Infine, vengono illustrate le principali misure di distanza o di similarità tra le unità statistiche, anch'esse diverse a seconda del tipo di variabili coinvolte (Paragrafo 3.5).

3.2 La matrice dei dati

Nella sua forma più comune, una matrice dei dati è una tabella contenente le informazioni disponibili relativamente ad un insieme di unità statistiche. In generale, supponendo di avere osservato i valori di *p* caratteri su un collettivo di *n* unità statistiche, la matrice dei dati, denotata ***X*** (di dimensione *n* x *p*), sarà strutturata nel modo seguente:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1h} & \dots & x_{1j} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2h} & \dots & x_{2j} & \dots & x_{2p} \\ \dots & & & & & & & \\ x_{i1} & x_{i2} & \dots & x_{ih} & \dots & x_{ij} & \dots & x_{ip} \\ \dots & & & & & & & \\ x_{r1} & x_{r2} & \dots & x_{rh} & \dots & x_{rj} & \dots & x_{rp} \\ \dots & & & & & & & \\ x_{n1} & x_{n2} & \dots & x_{nh} & \dots & x_{nj} & \dots & x_{np} \end{bmatrix}.$$

Ciascuna riga della matrice contiene le *p* informazioni relative ad una determinata unità statistica, mentre ciascuna colonna contiene le modalità assunte da un determinato carattere nelle diverse unità statistiche. Il suo generico elemento x_{ij} rappresenta dunque la modalità che il *j*-esimo carattere assume in corrispondenza della *i*-esima unità.
La matrice dei dati può anche essere vista come un insieme di *n* vettori riga (di dimensioni *1* x *p*) contenenti ciascuno il profilo di ciascuna unità statistica, ovvero i valori che in essa assumono le *p* variabili osservate.
Se si indica con $\mathbf{x}_i$ il generico vettore riga:

$$\mathbf{x}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \dots \\ x_{ip} \end{bmatrix}$$

la matrice **X** può dunque essere rappresentata nel modo seguente

$$\mathbf{X} = \left[\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_i, \dots \mathbf{x}_n\right]'.$$

Come accennato, i caratteri osservati possono essere di natura eterogenea. Si possono infatti avere: caratteri qualitativi sconnessi, o in scala nominale, che prevedono modalità non numeriche e non ordinabili; caratteri qualitativi ordinali, o in scala ordinale, che prevedono modalità non numeriche ma tra loro ordinabili; caratteri quantitativi, sia su scala ad intervalli che su scala di rapporti, che prevedono modalità numeriche. La matrice dei dati è pertanto spesso caratterizza dalla presenza di variabili miste, alcune quantitative e altre qualitative.
Come le variabili, anche le unità statistiche possono essere di varia natura. Nelle analisi intra-aziendali, ad esempio, le unità osservate possono essere i singoli prodotti dell'azienda, oppure i diversi stabilimenti produttivi, i diversi reparti di uno stabilimento, i dipendenti o i clienti dell'azienda. Nelle analisi inter-aziendali, invece, le unità osservate possono essere costituite dalle

diverse aziende concorrenti o da un campione di consumatori dei prodotti del settore. In altre analisi si possono avere, quali unità statistiche, le diverse regioni di un paese o i diversi settori produttivi.

3.3 La qualità dei dati e le mancate risposte parziali

Una volta costruita la matrice dei dati, un problema da affrontare in via preliminare riguarda la qualità delle informazioni in essa contenute. Come si è visto nei due precedenti capitoli, problemi di qualità possono riguardare sia i dati derivanti da rilevazioni primarie che quelli desunti da fonti secondarie. Nel primo caso alcuni problemi di qualità possono essere prevenuti nella fase della rilevazione: attraverso la formulazione chiara ed univoca dei quesiti; la scelta di una adeguata tecnica di acquisizione delle informazioni; l'addestramento dei rilevatori; la realizzazione di indagini pilota, ecc. Nel secondo caso si tratta invece di acquisire, insieme ai dati, anche le informazioni sul loro processo di formazione e sulla loro qualità, al fine di valutarne l'utilizzabilità nello specifico contesto di analisi.
I due principali problemi di qualità in una matrice dei dati sono costituiti dalla presenza di valori errati o di valori mancanti. I primi possono derivare da diverse, come si vedrà nel prossimo paragrafo, e occorre in primo luogo individuarli. I secondi sono invece direttamente individuabili, derivando o da mancate risposte parziali nelle rilevazioni primarie o da "buchi informativi" presenti nelle fonti statistiche di dati secondari, e il problema è scegliere il modo di trattarli, come si vedrà nel paragrafo successivo.

3.3.1 I valori errati

Nel caso di dati rilevati tramite indagine i valori errati possono derivare da:

- errore di risposta dell'intervistato che non ha interpretato correttamente il quesito posto, ha voluto deliberatamente rispondere in modo errato o, infine, ha fornito una risposta imprecisa per problemi di memoria o non conoscenza puntuale del fenomeno;
- errore dell'intervistatore nel porre il quesito;
- errore nel *data entry* o negli strumenti automatici di acquisizione su supporto informatico dei questionari cartacei[1].

La presenza di valori errati può essere segnalata da:

- valori fuori dominio, cioè non appartenenti ad un insieme predefinito di valori ammissibili;
- valori anomali, o *outlier*, cioè significativamente diversi da quelli osservati nella maggior parte delle unita;
- incompatibilità di risposte all'interno dello stesso questionario, quando i valori di una o piu variabili in esso rilevate contraddicono predefinite regole di natura logica e/o relazioni di tipo matematico.

L'individuazione dei valori errati in genere si avvale di una serie di controlli, classificabili nelle seguenti categorie:

- *controlli di consistenza*. Verificano che prefissate combinazioni di valori assunti da variabili rilevate in una stessa unità soddisfino determinati requisiti (regole di incompatibilità);
- *controlli di validità o di range*. Verificano che i valori assunti da una data variabile siano interni all'intervallo di definizione della variabile stessa;
- *controlli per gli outlier*. Utilizzati per isolare le unità statistiche che presentano, per alcune delle variabili, valori che si discostano in modo significativo dai valori che le stesse assumono nel resto delle unità rilevate o rispetto a rilevazioni precedenti. Questi valori sono

[1] Nonostante il processo di informatizzazione metta a disposizione strumenti portatili di dimensioni sempre più compatte, il questionario cartaceo è sempre molto utilizzato specialmente per i questionari autosomministrati in luoghi aperti al pubblico, nei questionari postali, ecc. (vedi Capitolo 2).

con alta probabilità errati, ma l'asserzione della loro non correttezza necessita di ulteriori verifiche come, ad esempio, la reintervista dell'unità utilizzando un diverso strumento di rilevazione[2].

Gli errori possono essere dovuti ad una qualunque delle fasi di raccolta e messa a punto dei dati o ad una serie di concause. Per questo motivo, mentre tradizionalmente il processo di controllo e correzione avveniva in un momento successivo alla fase di registrazione dei dati, la tendenza attuale e quella di spostare il controllo dei dati il più possibile vicino alla fase di raccolta delle informazioni presso le unità statistiche, in modo da rendere possibile la correzione immediata di informazioni che risultassero non compatibili o anomale.

3.3.2 Le mancate risposte parziali

Le mancate risposte a una o più domande nelle rilevazioni campionarie, o l'assenza di qualche dato nelle fonti secondarie, possono essere trattate in diversi modi. Un primo e più sbrigativo modo consiste nella eliminazione dalla matrice dei dati di tutte le unità osservate solo parzialmente. L'utilizzo delle sole unità osservate completamente determina però a sua volta due non trascurabili conseguenze negative. La prima è la riduzione della numerosità delle osservazioni, che influenza l'efficienza delle stime. La seconda conseguenza, forse ancor più grave della prima, si può determinare sulla correttezza delle stime. Ciò avviene quando la presenza di dati mancanti non è casuale, ovvero quando la probabilità di osservare una mancata risposta parziale per una variabile dipende da altre caratteristiche delle unità.
Una diversa soluzione al problema, frequentemente adottata, consiste nell'utilizzare diversi insiemi di unità per realizzare le diverse analisi dei dati. Si scartano cioè non tutte le unità osservate parzialmente, ma solo quelle con dati mancanti relativamente alle variabili di interesse per ogni specifica analisi. Le analisi univariate saranno cioè basate solo sulle risposte effettivamente fornite per quella specifica domanda; le analisi bivariate solo sulle unità che hanno i dati completi per quella coppia di variabili; le analisi multivariate saranno basate solo sulle unità che presentano risposte valide per tutte le variabili utilizzate nella specifica analisi.
Adottando questa soluzione, la conseguenza negativa è che la numerosità del campione risulterà variabile nelle diverse analisi effettuate, il che complica la lettura e l'interpretazione dei risultati.
Una soluzione ancora diversa, per molti aspetti preferibile a tutte le precedenti, consiste nell'imputare la mancata risposta, assegnando al dato mancante un valore 'plausibile'. Ciò consente di mantenere la dimensione originaria della matrice per tutte le analisi da realizzare.
Le tecniche di imputazione possono essere diverse. Le più comuni sono le seguenti.
Imputazione di un valore medio. Alla mancata risposta parziale viene sostituito il valore medio calcolato sulle unità osservate. In genere si utilizza la media aritmetica o la mediana per caratteri quantitativi o qualitativi ordinali e la moda per caratteri qualitativi sconnessi. Se si utilizza la media aritmetica, una volta effettuata l'imputazione non cambia la media della distribuzione ma si riduce la variabilità poiché tutti i dati imputati assumono lo stesso valore. Se invece si utilizza la mediana, si lascia invariata la mediana stessa, si avvicinano media e mediana e si riduce la variabilità.
Una modalità un po' più complessa di applicazione di tale metodo consiste nell'imputare la mancata risposta parziale con la media o la mediana calcolate, invece che sul complesso delle unità, su un sottoinsieme più omogeneo rispetto all'unità oggetto di imputazione. In altri termini, per una determinata variabile con mancate risposte, si suddivide l'insieme delle unità osservate in sottoinsiemi omogenei - possibilmente costruiti sulla base di una o più variabili prive di mancate risposte e legate alla variabile oggetto di imputazione - e si imputano i dati mancanti di ogni sottoinsieme con la rispettiva media (o mediana). Ne deriva una modificazione della media stimata prima delle imputazioni, in direzione di quella vera della popolazione, e un minore effetto di

[2] Si tratta, ad esempio, delle reinterviste telefoniche o dirette sui questionari 'dubbi' a seguito di un'indagine postale.

appiattimento sugli indici di variabilità.

Imputazione con prelievo da donatore. Invece che un valore medio, a ciascuna mancata risposta parziale viene imputato un valore individuale "donato" da una unità il più possibile simile a quella con dato mancante. Per individuare l'unità donatore di risposta i metodi più utilizzati sono quelli denominati *cold deck* e *hot deck*. Entrambi prevedono le seguenti fasi:

- si dividono le unità in due gruppi: quelli con mancate risposte parziali, che necessitano di ricevere un dato, e quelli completi, che costituiscono i potenziali donatori;
- a ciascuna mancata risposta parziale si imputa la risposta data dall'unità più simile in relazione ad altre caratteristiche presenti nella matrice dei dati; qualora si individuino più unità come potenziali donatori se ne sceglie una casualmente.

La differenza tra i due metodi sta nel fatto che, mentre con il metodo *cold deck* il gruppo dei potenziali donatori resta il medesimo per tutto il processo di imputazione, con il metodo *hot deck* esso viene aggiornato ad ogni imputazione successiva.

Imputazione da modello. Per ciascuna variabile affetta da osservazioni mancanti si specifica un modello di regressione multipla utilizzando k regressori scelti tra le altre variabili presenti nella matrice dei dati, in qualche modo collegate alla variabile da imputare. Effettuata la stima dei parametri del modello, a ciascuna unità con mancata risposta si imputa il valore predetto dal modello, dati i valori in essa assunti dai regressori. Nel caso l'unità considerata presenti mancate risposte anche tra le variabili esplicative del modello in genere si utilizza un modello con un minore numero di regressori.

Imputazione stocastica. A ciascun dato mancante se ne sostituisce uno estratto casualmente da una distribuzione ritenuta plausibile per la variabile. Il problema principale consiste nell'individuare, di volta in volta, la più plausibile distribuzione delle variabili affette da mancate risposte. Il vantaggio sta invece nella possibilità di ripetere più volte la procedura e ricavare utili indicazioni sulla variabilità introdotta nelle stime dal processo di imputazione.

3.4 Le analisi sui profili di colonna

Come accennato nell'introduzione, le analisi sui profili di colonna della matrice di dati vengono di seguito limitate a quelle bivariate, volte ad analizzare le associazioni esistenti tra le variabili considerate a coppie. In generale, l'obiettivo delle analisi bivariate è ottenere, a partire dalla matrice dei dati **X** (di dimensioni n x p) una matrice delle associazioni **A** (di dimensioni p x p) del tipo seguente:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1h} & \dots & a_{1j} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2h} & \dots & a_{2j} & \dots & a_{2p} \\ \dots & & & & & & & \\ a_{h1} & a_{h2} & \dots & a_{hh} & \dots & a_{hj} & \dots & a_{hp} \\ \dots & & & & & & & \\ a_{j1} & a_{j2} & \dots & a_{jh} & \dots & a_{jj} & \dots & a_{jp} \\ \dots & & & & & & & \\ a_{p1} & a_{p2} & \dots & a_{ph} & \dots & a_{pj} & \dots & a_{pp} \end{bmatrix},$$

dove a_{hj} è una misura dell'associazione esistente tra la h-esima e la j-esima variabile della matrice dei dati.
La misura di tale associazione dipende dal tipo di variabili considerate, che come più volte ricordato possono essere qualitative sconnesse, qualitative ordinali, quantitative, miste. Di seguito vengono richiamati i principali indici di associazione utilizzabili in presenza di queste diverse tipologie di variabili.

Variabili qualitative sconnesse. A partire da ciascuna coppia di colonne della matrice dei dati può essere costruita una distribuzione doppia di frequenze e su di essa si può calcolare l'**indice di associazione chi-quadrato**, definito come:

$$\chi^2 = \sum_{s=1}^{S} \sum_{t=1}^{T} \frac{c_{st}^{\ 2}}{n'_{st}},$$

dove:
S e T sono rispettivamente il numero di modalità del primo e del secondo carattere;
$n'_{st} = (n_{s.}\ n_{.t}) / n$ sono le frequenze teoriche di indipendenza nella tabella doppia, cioè le frequenze che si sarebbero osservate, dati i valori marginali di riga e di colonna, nel caso di perfetta indipendenza tra i due caratteri;
$c_{st} = n_{st} - n'_{st}$ sono le contingenze, cioè le differenze tra le frequenze osservate e quelle teoriche di indipendenza.
Se i due caratteri sono perfettamente indipendenti, tutti i numeratori dei rapporti sono pari a zero e quindi l'indice assume valore zero. Quanto più le frequenze osservate si discostano dalle frequenze teoriche di indipendenza l'indice χ^2 assume valori più elevati. L'indice non ha un massimo definito e il suo valore, a parità di associazione, dipende dalla numerosità del collettivo.
Una misura relativa di associazione, che assume valori compresi tra zero ed uno, è data dall'**indice *v* di Cramér**, denominato indice medio di contingenza e definito come:

$$v = \left(\frac{\chi^2 / n}{\min[(S-1),(T-1)]} \right)^{1/2} .$$

L'indice v vale ancora zero quando i due caratteri sono indipendenti, mentre assume valore pari ad uno quando tra i due caratteri vi è massima associazione. Calcolato l'indice v per ogni coppia di variabili (colonne) della matrice dei dati originaria, v_{hj} per le generiche colonne h-esima e j-esima, si otterrà dunque una matrice V (di dimensioni p x p) del tipo seguente:

$$V = \begin{bmatrix} v_{11} & v_{12} & \dots & v_{1h} & \dots & v_{1j} & \dots & v_{1p} \\ v_{21} & v_{22} & \dots & v_{2h} & \dots & v_{2j} & \dots & v_{2p} \\ \dots & & & & & & & \\ v_{h1} & v_{h2} & \dots & v_{hh} & \dots & v_{hj} & \dots & v_{hp} \\ \dots & & & & & & & \\ v_{j1} & v_{j2} & \dots & v_{jh} & \dots & v_{jj} & \dots & v_{jp} \\ \dots & & & & & & & \\ v_{p1} & v_{p2} & \dots & v_{ph} & \dots & v_{pj} & \dots & v_{pp} \end{bmatrix} .$$

La matrice V è simmetrica con valori unitari sulla diagonale principale.

Esempio 3.1
Supponiamo di disporre della seguente matrice dei dati relativa a tre variabili qualitative sconnesse - sesso, condizione lavorativa e sport preferito - rilevate su nove individui e di voler calcolare il grado di associazione tra le due variabili.

Individui	Sesso	Condizione	Sport
1	F	Impegato	Calcio
2	M	Operaio	Basket
3	F	Impiegato	Calcio
4	M	Libero prof.	Nuoto
5	M	Impiegato	Calcio
6	M	Operaio	Calcio
7	F	Operaio	Nuoto
8	M	Operaio	Nuoto
9	F	Impiegato	Basket

Tra la coppia di caratteri sesso e condizione lavorativa si può costruire la seguente tabella a doppia entrata

Sesso	Condizione lavorativa			
	Impiegato	Lib. Prof.	Operaio	Totale
F	3	0	1	4
M	1	1	3	5
Totale	4	1	4	9

Da essa si possono ricavare le seguenti tabelle delle frequenze teoriche e delle contingenze:

Frequenze teoriche:

Sesso	Condizione lavorativa			
	Impiegato	Lib. Prof.	Operaio	Totale
F	1,78	0,44	1,78	4

M	2,22	0,56	2,22	5
Totale	4	1	4	9

Contingenze:

Sesso	Condizione lavorativa			
	Impiegato	Lib. Prof.	Operaio	Totale
F	1,22	-0,44	-0,78	0
M	-1,22	0,44	0,78	0
Totale	0	0	0	0

L'indice di associazione chi-quadrato è pari a:

$$\chi^2 = \frac{1,488}{1,78} + \frac{0,194}{0,44} + \frac{0,601}{1,78} + \frac{1,488}{2,22} + \frac{0,194}{0,56} + \frac{0,601}{2,22} = 2,9 ;$$

mentre l'indice v è pari a:

$$v = \sqrt{\frac{2,9/9}{1}} = 0,57 .$$

Procedendo in modo analogo per le altre coppie di caratteri si ottiene la seguente matrice di associazione V ;

$$V = \begin{vmatrix} 1 & 0,57 & 0,16 \\ 0,57 & 1 & 0,52 \\ 0,16 & 0,52 & 1 \end{vmatrix} .$$

Variabili qualitative ordinali. Nel caso di variabili qualitative ordinali un indice di associazione deve poter misurare, oltre all'intensità dell'associazione presente tra i due caratteri, anche il verso della relazione. L'indice deve cioè poter distinguere tra una relazione positiva, dove al crescere delle modalità di un carattere tendono a crescere anche le modalità dell'altro, da una relazione negativa, dove al crescere delle modalità di un carattere quelle dell'altro tendono a decrescere.
Se il numero delle modalità dei due caratteri non è troppo elevato, si può calcolare, a partire dalla distribuzione doppia di frequenze, l'**indice gamma di Goodman e Kruskal**:

$$\gamma = \frac{N_c - N_d}{N_c + N_d} ,$$

dove N_c e N_d sono, rispettivamente, il numero di coppie in cui i caratteri sono ordinati allo stesso modo, e manifestano quindi una concordanza, e il numero di coppie in cui, al contrario, i due caratteri sono ordinati in modo diverso (discordanza). Una coppia di unità evidenzia concordanza, e pertanto appartiene a N_c quando, rispetto alla prima unità, le modalità che i due caratteri assumono nella seconda sono entrambe maggiori o entrambe minori. Al contrario, una coppia di unità evidenzia discordanza, e dunque appartiene a N_d, quando, rispetto alla prima unità, le modalità che i due caratteri assumono nella seconda sono una maggiore e l'altra minore.
L'indice γ assume valori compresi tra –1 e 1, dove il segno indica il verso della relazione (concordanza o discordanza), mentre il valore assoluto indica l'intensità della stessa. L'indice assume infatti il suo valore massimo quando $N_d = 0$, quando cioè tutte le coppie di unità evidenziano concordanza, mentre assume valore pari a –1 quando $N_c = 0$ e quindi tutte le coppie esprimono discordanza.

Se invece il numero delle modalità dei due caratteri è elevato, avvicinandosi al numero delle unità del collettivo, si può utilizzare l'indice di associazione tra graduatorie **ρ di Spearman** definito come:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^{\,2}}{n(n^2-1)},$$

dove d_i indica la differenza tra le posizioni in graduatoria per i due caratteri in esame relativa alla i-esima unità.
Calcolati γ o ρ per tutte le coppie di caratteri, si otterranno, rispettivamente, le matrici Γ o P (entrambe di dimensioni p x p).

Esempio 3.2
Si dispone dei seguenti dati relativi ai livelli di soddisfazione di 13 clienti di un supermercato per i reparti taglio, macelleria e pescheria.

Clienti	Reparto taglio	Reparto macelleria	Reparto pescheria
1	alto	medio	medio
2	medio	alto	basso
3	alto	alto	basso
4	basso	medio	medio
5	basso	basso	medio
6	medio	basso	alto
7	alto	medio	alto
8	medio	alto	basso
9	alto	alto	medio
10	basso	medio	alto
11	alto	alto	alto
12	alto	basso	medio
13	medio	medio	basso

Per i caratteri soddisfazione per il reparto taglio e soddisfazione per il reparto macelleria si può costruire la seguente tabella a doppia entrata

Soddisfazione reparto taglio	Soddisfazione reparto macelleria			
	Alto	Medio	Basso	Totale
Alto	3	2	1	6
Medio	2	1	1	4
Basso	0	2	1	3
Totale	5	5	3	13

Partendo dall'unica unità che presenta basso livello di soddisfazione sia per il reparto taglio che per il reparto macelleria, le altre unità che esprimono concordanza con questa sono: le tre unità che esprimono alto livello di soddisfazione per entrambi i caratteri; le due unità che esprimono alta soddisfazione per il reparto taglio e media per il reparto macelleria; le due unità che esprimono media soddisfazione per il reparto taglio e alta per il reparto macelleria, oltre all'unità che esprime media soddisfazione per entrambi i reparti.
Il numero di coppie di unità che esprimono concordanza e che comprendono il cliente con bassa soddisfazione per entrambi i caratteri e pari a 1(3+2+2+1). Il numero complessivo di coppie di unità che esprimono concordanza è dunque pari a:

N_c=1(3+2+2+1)+1(3+2)+1(0)+2(3+2)+1(3)+2(0)+0(0)+2(0)+3(0)=26 .

Partendo dal vertice della tabella che contiene il numero di clienti con bassa soddisfazione per il reparto taglio e alta soddisfazione per il reparto macelleria si può calcolare il numero di coppie che esprimono discordanza nel modo seguente:

N_d = 0(2 + 1 + 1 + 1) + 2(2 + 1) + 2(1 + 1) + 1(1) = 11 ;

L'indice γ risulta pertanto pari a:

$$\gamma = \frac{26-11}{26+11} = \frac{15}{37} = 0{,}41 \ .$$

Procedendo in modo analogo per le altre coppie di variabili si ottiene la matrice di associazione

$$\Gamma = \begin{vmatrix} 1 & 0{,}41 & 0{,}05 \\ 0{,}41 & 1 & -0{,}49 \\ 0{,}05 & -0{,}49 & 1 \end{vmatrix} .$$

Esempio 3.3
Si ipotizzi che ad un campione di consumatori siano state fatte assaggiare cinque marche di pasta con il medesimo condimento e che siano stati rilevati gli ordinamenti in relazione alle caratteristiche gusto e capacità di tenere la cottura, attribuendo la prima posizione (1) alla pasta giudicata migliore e l'ultima (5) alla peggiore. Si ipotizzi inoltre che con una rilevazione presso un campione di esercizi commerciali siano stati rilevati i prezzi medi di vendita delle cinque marche di pasta, espressi anch'essi su scala ordinale, assegnando punteggio pari ad 1 alla marca più costosa e così via fino al punteggio 5 assegnato alla più economica. Si supponga che la matrice dei dati sia la seguente:

Marca di pasta	Gusto	Cottura	Prezzo
A	1	3	3
B	2	1	4
C	4	5	5
D	3	2	2
E	5	4	1

Tra i caratteri gusto e capacità di tenere la cottura si calcolano le differenze tra le posizioni in graduatoria e le differenze al quadrato:

Marca di pasta	Posizioni in graduatoria		d	d^2
	Gusto	Cottura		
A	1	3	-2	4
B	2	1	1	1
C	4	5	-1	1
D	3	2	1	1
E	5	4	1	1

L'indice ρ sarà pertanto pari a:

$$\rho = 1 - \frac{6 \cdot 8}{5(25-1)} = 1 - \frac{48}{120} = 0{,}60 \ .$$

Procedendo in modo analogo per le altre coppie di variabili si ottiene la matrice di associazione seguente:

$$\mathrm{P} = \begin{vmatrix} 1 & 0{,}6 & 0{,}4 \\ 0{,}6 & 1 & -0{,}3 \\ 0{,}4 & -0{,}3 & 1 \end{vmatrix} .$$

Caratteri quantitativi. Le più comuni misure di associazione tra variabili quantitative sono la covarianza e il coefficiente di correlazione lineare. Indicati con x_{ih} e x_{ij} i valori assunti dalle variabili *h*-esima e *j*-esima nella unità *i*-esima e con $\bar{x}_h$ e $\bar{x}_j$ i rispettivi valori medi, la **covarianza** è definita dalla seguente espressione:

$$s_{hj} = \frac{\sum_{i=1}^{n} (x_{ih} - \bar{x}_h)(x_{ij} - \bar{x}_j)}{n}$$

Si hanno valori positivi della covarianza quando la somma algebrica dei prodotti al numeratore è positiva e quindi prevalgono prodotti di segno positivo, che indicano concordanza tra le due variabili (valori di x_h e x_j congiuntamente o maggiori delle rispettive medie aritmetiche, o minori di esse). Valori negativi si hanno invece quando prevalgono prodotti che indicano discordanza tra i due caratteri (valori di x_h maggiori della media e valori di x_j minori, o viceversa).
Valori pari a zero si hanno infine quando la somma algebrica dei prodotti al numeratore si annulla, cioè quando non prevalgono né i prodotti che indicano concordanza né quelli che indicano discordanza, ovvero i due caratteri sono linearmente indipendenti.
Calcolata la covarianza per tutte le coppie di variabili si può costruire una matrice del tipo seguente (di dimensioni *p* x *p*), detta matrice delle covarianze:

$$S = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1h} & \dots & s_{1j} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2h} & \dots & s_{2j} & \dots & s_{2p} \\ \dots & & & & & & & \\ s_{h1} & s_{h2} & \dots & s_{hh} & \dots & s_{hj} & \dots & s_{hp} \\ \dots & & & & & & & \\ s_{j1} & s_{j2} & \dots & s_{jh} & \dots & s_{jj} & \dots & s_{jp} \\ \dots & & & & & & & \\ s_{p1} & s_{p2} & \dots & s_{ph} & \dots & s_{pj} & \dots & s_{pp} \end{bmatrix} .$$

La matrice è simmetrica e sulla diagonale principale presenta le varianze delle *p* variabili, ovvero le covarianze di ogni variabile con se stessa. Per la generica variabile *h* si ha infatti:

$$s_{hh} = \frac{\sum_{i=1}^{n} (x_{ih} - \bar{x}_h)(x_{ih} - \bar{x}_h)}{n} = \frac{\sum_{i=1}^{n} (x_{ih} - \bar{x}_h)^2}{n} = \sigma^2{}_h$$

I valori assunti dalle covarianze dipendono dalle scale di misura delle variabili nella matrice dei dati, il che li rende non direttamente confrontabili tra loro al fine di valutare se tra una coppia di variabili vi sia una associazione maggiore o minore rispetto ad un'altra.
Per ovviare a tale problema si può ricorrere al **coefficiente di correlazione lineare di Bravais e Pearson**, definito dalla seguente espressione:

$$r_{hj} = \frac{s_{hj}}{\sigma_h \quad \sigma_j},$$

dove σ_h e σ_j sono le deviazioni standard dei due caratteri. Il coefficiente di correlazione lineare assume lo stesso segno della covarianza ed è compreso tra –1 e 1.
Calcolati i coefficienti di correlazione lineare tra tutte le coppie di variabili si ottiene dunque una matrice di correlazione **R**, del tipo seguente:

$$R = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1h} & \dots & r_{1j} & \dots & r_{1p} \\ r_{21} & r_{22} & \dots & r_{2h} & \dots & r_{2j} & \dots & r_{2p} \\ \dots & & & & & & & \\ r_{h1} & r_{h2} & \dots & r_{hh} & \dots & r_{hj} & \dots & r_{hp} \\ \dots & & & & & & & \\ r_{j1} & r_{j2} & \dots & r_{jh} & \dots & r_{jj} & \dots & r_{jp} \\ \dots & & & & & & & \\ r_{p1} & r_{p2} & \dots & r_{ph} & \dots & r_{pj} & \dots & r_{pp} \end{bmatrix}.$$

La matrice è simmetrica e con valori unitari sulla diagonale principale.

Esempio 3.4
Si ipotizzi di aver rilevato alcuni indici di bilancio su quattro aziende, riportati nella seguente matrice dei dati.

Azienda	ROI	ROE	ROS	Indebitamento
A	7,1	8,2	5,7	25,3
B	0,3	1,5	2,8	53,4
C	0,4	2,5	5,2	34,1
D	5,1	2,0	1,2	11,7

Relativamente agli indici ROI e ROE si calcolano medie aritmetiche e deviazioni standard:
$\bar{x}_{ROI} = 3{,}23$; $\bar{x}_{ROE} = 3{,}55$; $\sigma_{ROI} = 2{,}96$; $\sigma_{ROE} = 2{,}71$.

Gli indici di covarianza e di correlazione si calcolano nel modo seguente:

$$s_{ROIROE} = \frac{(7{,}1-3{,}23)(8{,}2-3{,}55)+(0{,}3-3{,}23)(1{,}5-3{,}55)+(0{,}4-3{,}23)(2{,}5-3{,}55)+(5{,}1-3{,}23)(2{,}0-3{,}55)}{5} =$$

$$= \frac{24{,}08}{4} = 6{,}02 \ ;$$

$$r_{ROIROE} = \frac{6{,}02}{2{,}96 * 2{,}71} = 0{,}75 \ .$$

Procedendo in modo analogo per tutte le altre coppie di variabili di ottengono le matrici ***S*** e ***R***:

$$S = \begin{bmatrix} 8,77 & 6,02 & 0,36 & -33,14 \\ 6,02 & 7,33 & 3,36 & -11,14 \\ 0,36 & 3,36 & 3,33 & 5,33 \\ -33,14 & -11,44 & 5,33 & 229,07 \end{bmatrix};$$

$$R = \begin{bmatrix} 1,00 & 0,75 & 0,07 & -0,74 \\ 0,75 & 1,00 & 0,68 & -0,28 \\ 0,07 & 0,68 & 1,00 & 0,19 \\ -0,74 & -0,28 & 0,19 & 1,00 \end{bmatrix}.$$

Caratteri misti. Come si è più volte ripetuto, nelle analisi statistiche di dati aziendali o di mercato la matrice dei dati è spesso a carattere misto, con alcune variabili quantitative e altre qualitative, ordinali o sconnesse.
Disponendo di una matrice di dati con variabili miste, occorre adottare misure di associazione diverse per ogni combinazione di tipologie di variabili da analizzare. Naturalmente, le misure di associazione relative a variabili dello stesso tipo (qualitative sconnesse, ordinali, quantitative) saranno quelle già illustrate nei punti precedenti, mentre problemi di scelta di indici appropriati si pongono nel caso in cui si debba misurare l'associazione tra variabili appartenenti a tipologie differenti. I possibili casi sono i seguenti:

a) - variabili qualitative sconnesse con qualitative ordinali;
b) - variabili qualitative sconnesse con variabili quantitative;
c) - variabili qualitative ordinali con variabili quantitative.

In questi casi, una prima soluzione consiste nel ricondurre le variabili alla medesima tipologia trasformando opportunamente la scala di misura delle variabili: nel caso a) riconducendo le variabili misurate su scala ordinale in sconnesse; nei casi b) e c) riconducendo le variabili misurate su una scala quantitativa in qualitativa (sconnessa nel primo caso; ordinale nel secondo). Ricondotte le variabili alla stessa tipologia, si calcolano i relativi indici già illustrati. Il limite di tale soluzione sta ovviamente nel fatto che viene perduta una parte dell'informazione contenuta nella matrice dei dati originari.
Un diverso approccio al problema, che supera l'inconveniente appena richiamato, consiste nel calcolare un indice di associazione introdotto appositamente per il caso di variabili miste. Tale indice è il **rapporto di correlazione**, dato dalla seguente espressione:

$$\eta^2_{y/x} = \frac{\sigma^2_{media(y/x)}}{\sigma^2_y},$$

dove:
x e *y* indicano, rispettivamente, la variabile qualitativa (sconnessa o ordinale) e quella quantitativa;
σ^2_y è la varianza della variabile quantitativa;
$\sigma^2_{media(y/x)}$ è la varianza delle medie del carattere quantitativo condizionate alle modalità di quello qualitativo.
Il rapporto di correlazione esprime dunque la quota della varianza complessiva di una variabile quantitativa spiegata dalle medie condizionate alle modalità di una variabile qualitativa: quando tali medie sono uguali tra loro vuol dire che la variabile *y* non dipende, in media, dalle modalità della variabile *x* e l'indice vale zero; quando tali medie sono invece diverse tra loro, vuol dire che esiste una relazione tra le modalità delle due variabili, che ha intensità massima quando tutta la varianza di *y* è spiegata dalla variabilità tra le medie condizionate, nel qual caso l'indice vale uno.

Calcolati i rapporti di correlazione tra tutte le coppie di variabili si ottiene dunque una matrice **E** di rapporti di correlazione analoga alle precedenti.

Il quadro completo delle misure di associazione utilizzabili nel caso di variabili miste è riportato nella matrice seguente, dove i diversi blocchi corrispondono ai possibili incroci tra le variabili delle diverse tipologie.

	Qualitativi sconnessi	Qualitativi ordinali	Quantitativi
Qualitativi sconnessi	**V**	**V**	**V** **E**
Qualitativi ordinali		**Γ, P**	**Γ, P** **E**
Quantitativi			**S e R**

Nei tre blocchi diagonali si possono calcolare le matrici **V**, **Γ, P** , **S** ed **R** viste in precedenza. Nei blocchi non diagonali, che si riferiscono all'associazione tra un carattere quantitativo e uno qualitativo, si possono calcolare alcune delle medesime matrici (**V**, **Γ, P)**, una volta ricondotte le variabili quantitative a sconnesse o a ordinali, a seconda dei casi, oppure si può calcolare la matrice dei rapporti di correlazione **E**.

Esempio 3.5
Si supponga di disporre delle seguenti informazioni relativamente ad un gruppo di clienti di una azienda commerciale: professione, sesso, livello di soddisfazione l'esercizio commerciale (LS), per il nostro esercizio, livello di soddisfazione per il principale *competitor* (LSC), reddito, quota acquisti presso l'esercizio.

Clienti	Professione	Sesso	LS	LSC	Reddito	Quota acquisti
1	Impiegato	F	Alto	Medio	1,8	58,1
2	Operaio	M	Medio	Basso	1,2	65,0
3	Impiegato	M	Basso	Basso	1,4	64,5
4	Operaio	F	Medio	Basso	1,3	63,0
5	Lav. Aut.	M	Alto	Alto	2,0	51,0
6	Impiegato	F	Alto	Basso	1,1	70,0
7	Operaio	M	Basso	Medio	1,5	40,0

8	Lav. Aut.	M	Medio	Basso	2,5	42,5
9	Operaio	M	Alto	Alto	1,4	66,0
10	Lav. Aut.	F	Medio	Basso	3,0	38,5
11	Operaio	M	Alto	Medio	1,0	60,0

Tra la professione e il reddito si può calcolare il rapporto di correlazione nel modo seguente:

$\text{media}_{(\text{reddito})} = 1{,}65$; $\sigma^2_{\text{reddito}} = 0{,}353$;

$\text{media}_{(\text{reddito/prof=impiegato})} = 1{,}43$; $\text{media}_{(\text{reddito/prof=operaio})} = 1{,}28$; $\text{media}_{(\text{reddito/prof=lav.autonomo})} = 2{,}50$.

$$\eta^2_{\textit{reddito / professioe}} = \frac{\frac{(1{,}43-1{,}65)^2 3 + (1{,}28-1{,}65)^2 5 + (2{,}50-1{,}65)^2 3}{11}}{0{,}353} = \frac{0{,}272}{0{,}353} = 0{,}77 .$$

La matrice di associazione è la seguente:

	Professione	Sesso	LS	LSC	Reddito	Quota acquisti
Professione	1,00	0,40	0,38	0,33	0,77	0,52
Sesso		1,00	0,41	0,39	0,03	0,01
LS			1,00	0,63	0,19	0,16
LSC				1,00	0,04	0,04
Reddito					1,00	-0,79
Quota acq.						1,00

Applicazioni in sas
Proc Corr

3.5 Le analisi sui profili di riga

Le analisi sui profili di riga hanno l'obiettivo di misurare in modo sintetico la "distanza" o la "similarità" tra coppie di unità del collettivo statistico, che sono appunto collocate nelle righe della matrice dei dati. La distanza ovviamente non è intesa in senso spaziale, bensì come differenza, tra le due unità, relativamente ai valori assunti dalle variabili contenute nella matrice dei dati. Ad esempio, ipotizziamo di avere condotto una indagine volta a misurare il livello di soddisfazione da parte dei clienti di una compagnia aerea in relazione a diverse caratteristiche (qualità dei servizi di terra, puntualità dei voli, cortesia del personale di volo, qualità del servizio di ristorazione, ecc). Misurare la distanza tra due clienti significa pervenire ad una misura sintetica di quanto essi sono tra loro diversi in relazione alle caratteristiche rilevate.
Indicata con d_{ir} la distanza tra le unità i-esima e r-esima, tale misura dovrebbe godere delle seguenti proprietà:
$d_{ir} \geq 0$, non negatività.Gli indici di distanza sono in genere positivi e risultano uguali a zero solo quando le due unità presentano le stesse modalità per tutti i caratteri presenti nella matrice dei dati;
$d_{ii} = 0$, la distanza tra ciascuna unità e se stessa è pari a zero;
$d_{ir} = d_{ri}$, simmetria. La distanza tra l'unità i-esima e l'unità r-esima deve risultare uguale a quella tra l'unità r-esima e la i-esima;
$d_{ir} \leq d_{is} + d_{sr}$, diseguaglianza triangolare. La somma delle distanze tra le unità *i* e *s* e le unità *s* e *r* deve essere minore o al più uguale alla distanza tra *i* e *r*.
Se una misura di distanza soddisfa tutte le proprietà sopra esposte si dice che lo spazio di riferimento è uno spazio metrico.
A partire dalla matrice dei dati $\boldsymbol{X}$, una volta calcolate tutte le distanze tra le n unità statistiche si ottiene una matrice di distanza $\boldsymbol{D}$, di dimensione n x n, del tipo seguente:

$$D = \begin{bmatrix} d_{11} & d_{12} & \dots & d_{1i} & \dots & d_{1r} & \dots & d_{1n} \\ d_{21} & d_{22} & \dots & d_{2i} & \dots & d_{2r} & \dots & d_{2n} \\ \dots & & & & & & & \\ d_{i1} & d_{i2} & \dots & d_{ii} & \dots & d_{ir} & \dots & d_{in} \\ \dots & & & & & & & \\ d_{r1} & d_{r2} & \dots & d_{ri} & \dots & d_{rr} & \dots & d_{rn} \\ \dots & & & & & & & \\ d_{n1} & d_{n2} & \dots & d_{ni} & \dots & d_{nr} & \dots & d_{nn} \end{bmatrix}$$

Nel seguito verranno illustrate le diverse misure di distanza che possono essere calcolate in corrispondenza alle diverse tipologie di caratteri presenti nella matrice dei dati.

Caratteri qualitativi sconnessi politomici. In presenza di una matrice dei dati composta da p caratteri qualitativi sconnessi politomici si può utilizzare l'indice di distanza di Sneath, dato dalla frequenza relativa dei caratteri per i quali le unità i-esima e r-esima presentano modalità diverse.
Per il generico carattere *k* si pone:

$d_{ir,k} =$ se $x_{ik} \neq x_{rk}$,
$d_{ir,k} = 0$ se $x_{ik} = x_{rk}$.

L'**indice di distanza di Sneath** tra *i* ed *r* è:

$$d_{ir} = \frac{\sum_{k=1}^{p} d_{ir,k}}{p}.$$

In sostanza si contano le colonne per le quali le righe *i* e *r* della matrice dei dati presentano modalità diverse e si rapporta tale numero a *p*, cioè al totale dei caratteri presenti nella matrice dei dati. Essendo definito come frequenza relativa, l'indice è compreso tra 0 e 1. Alternativamente, si può definire una corrispondente misura di similarità tra due unità, denotato c_{ir}, come la frequenza relativa dei caratteri con modalità uguali. Ovviamente sarà $d_{ir} + c_{ir} = 1$.
Nel seguito, per ogni misura di distanza comprese tra 0 e 1 verrà definito un corrispondente **indice di similarità** c_{ir} dato dal complemento a 1 dell'indice di distanza:

$c_{ir} = 1 - d_{ir}$.

Esempio 3.6
Si ipotizzi di voler misurare la distanza tra i soggetti a partire dalla seguente matrice dei dati contenente le marche preferite di caffè, thè e pasta, nonché l'insegna della grande distribuzione preferita in un campione di 5 soggetti:

Soggetto	Caffè	Thè	Pasta	Insegna
1	Splendid	Lipton	De Cecco	Conad
2	Sao	Infrè	Barilla	Sidis
3	Lavazza	Lipton	De Cecco	Conad
4	Splendid	Infrè	Barilla	Coop
5	Sao	Lipton	Barilla	Coop

La distanza tra i soggetti 1 e 2 è pari ad 1 (massimo) poiché in essi tutte le modalità dei quattro caratteri sono diverse. Si ha cioè: $d_{12,1} = 1$; $d_{12,2} = 1$; $d_{12,3} = 1$; $d_{12,4} = 1$.
L'indice di distanza sarà dunque:

$$d_{12} = \frac{1+1+1+1}{4} = 1 .$$

Invece, la distanza tra il soggetto 1 e 3 è molto minore poiché in tre casi su quattro le modalità coincidono. Si ha infatti: $d_{13,1} = 1$; $d_{13,2} = d_{13,3} = d_{13,4} = 0$ e quindi:

$d_{13} = 1/4 = 0{,}25$.

Calcolata per tutte le coppie di soggetti, la matrice delle distanze risulterà la seguente:

$$D = \begin{bmatrix} 0 & 1 & 0{,}25 & 0{,}75 & 0{,}75 \\ 1 & 0 & 1 & 0{,}5 & 0{,}5 \\ 0{,}25 & 1 & 0 & 1 & 0{,}751 \\ 0{,}75 & 0{,}5 & 1 & 0 & 0{,}5 \\ 0{,}75 & 0{,}5 & 0{,}75 & 0{,}5 & 0 \end{bmatrix}.$$

Caratteri qualitativi dicotomici. Ipotizziamo che la matrice dei dati contenga *p* misurazioni nominali dicotomiche del tipo presenza/assenza, indicate rispettivamente con 1 e 0. Due generiche righe della matrice dei dati possono essere sintetizzate nella seguente tabella di contingenza:

		unità i	
		1	0
unità r	1	*a*	*b*

0 c d

dove a rappresenta il numero di caratteri presenti in entrambe le unità, b il numero di caratteri presenti nell'unità r ma assenti nell'unità i, c il numero di caratteri presenti nell'unità i ma assenti nell'unità r e d il numero di caratteri assenti in entrambe le unità.
A partire da tali elementi possono essere calcolate diverse misure di distanza, secondo gli approcci di seguito richiamati.

Simple matching. La più immediata misura di distanza è data dalla frequenza relativa degli attributi presenti in una unità e assenti nell'altra:

$$d_{ir} = \frac{b+c}{p} .$$

Il corrispondente indice di similarità sarà:

$$c_{ir} = 1 - d_{ir} = \frac{a+d}{p} .$$

Indice di distanza di Jaccard. A differenza dell'approccio *simple matching*, l'indice di Jaccard esclude dal denominatore l'elemento d, cioè il numero di caratteri assenti in entrambe le unità. L'indice è dunque il seguente:

$$d_{ir} = \frac{b+c}{a+b+c} ,$$

mentre il corrispondente indice di similarità è:

$$c_{ir} = 1 - d_{ir} = \frac{a}{a+b+c} .$$

Questa diversa definizione del denominatore è suggerita dalla non opportunità di considerare come indicatore di similitudine tra due unità il fatto di presentare entrambe modalità assente in relazione ad un carattere. Per chiarire meglio si consideri, ad esempio, il caso di un sondaggio in cui si chiede ad alcuni individui se il sabato sera sono soliti andare in pizzeria. Considerando due unità, se entrambe rispondono sì ciò è sicuramente indice di similitudine, mentre se una risponde sì e l'altra no è altrettanto inequivocabilmente indice di diversità. Ma se entrambe rispondono no ciò non può essere considerato indice di similitudine perché un individuo potrebbe essere solito andare al pub, mentre l'altro potrebbe usualmente restare a casa.

Indice di distanza di Czekanowski. L'indice fa propria la critica di Jaccard all'approccio *Simple matching* e, in più, assegna un peso doppio al numero di attributi presenti in entrambe le unità. L'indice di distanza è dato pertanto dall'espressione:

$$d_{ir} = \frac{b+c}{2a+b+c} ,$$

mentre quello di similarità è:

$$c_{ir} = 1 - d_{ir} = \frac{2a}{2a+b+c}$$

Esempio 3.7
Si ipotizzi di aver rilevato i paesi in cui un gruppo di aziende concorrenti esportano i loro prodotti, ottenendo i seguenti dati:

Azienda	Francia	Germania	Paesi Bassi	Spagna	Austria	Grecia
A	Si	Si	No	Si	No	No

B	Si	No	No	Si	Si	Si
C	No	No	No	No	Si	No
D	Si	No	Si	Si	Si	Si

Se consideriamo la coppia di aziende BD avremo:

$a = 4$, $b = 1$, $c = 0$, $d = 1$, con $p = 6$.
Se si utilizza l'approccio *simple matching* avremo pertanto $d_{BD} = 1/6$. Utilizzando l'indice di Jaccard avremo $d_{BD} = 1/5$. Se si utilizza infine l'indice di Czekanowski si ottiene $d_{BD}=1/9$.
Per il solo indice di *simple matching*, la matrice completa delle distanze tra le quattro aziende è la seguente:

$$D = \begin{bmatrix} 0 & 1/6 & 4/6 & 4/6 \\ 1/6 & 0 & 3/6 & 1/6 \\ 4/6 & 3/6 & 0 & 4/6 \\ 4/6 & 1/6 & 4/6 & 0 \end{bmatrix}.$$

Caratteri qualitativi ordinali. Nel caso di una matrice dei dati contenente tutte variabili ordinali, come ad esempio il livello di soddisfazione per una serie di caratteristiche di un prodotto o servizio (in genere è rilevato su scala ordinale, con modalità del tipo: per niente; poco; abbastanza; molto; moltissimo) una possibile soluzione consiste nell'attribuire un punteggio crescente di una stessa quantità al crescere della misurazione ordinale. Ad esempio: 1 = per niente; 2 = poco; 3 = abbastanza; 4 = molto; 5 = moltissimo. Compiuta questa operazione si può utilizzare un indice di distanza per dati quantitativi (vedi più avanti).
Ovviamente, così facendo si introduce un elemento di arbitrarietà nell'analisi, poiché si ipotizza: a) che la differenza tra due modalità contigue sia sempre la stessa (ad esempio, che la differenza di soddisfazione che esiste tra poco e abbastanza sia pari a quella che esistente tra molto e moltissimo); b) che la differenza tra due modalità separate da una terza sia doppia rispetto a quella tra modalità contigue (ad esempio, che la differenza di soddisfazione tra abbastanza e moltissimo sia doppia rispetto a quella tra poco e abbastanza). L'alternativa consiste nel considerare la misurazione a livello nominale e applicare l'indice di Sneath; in questo modo però si rinuncia ad una gran parte del contenuto informativo presente nella matrice dei dati, ragion per cui in genere si preferisce adottare la prima soluzione.

Caratteri quantitativi. Nel caso di caratteri quantitativi la prima e più elementare misura di distanza che si può adottare è costituita dalla distanza euclidea, definita come la radice quadrata della somma delle differenze al quadrato tra le modalità appartenenti alle due unità relative a tutti i caratteri presenti nella matrice dei dati.
Indicate con x_{ik} e x_{rk} le modalità assunte dalla variabile k nelle unità i ed r, la distanza euclidea d_{ir} tra le due unità è data dalla seguente espressione:

$$d_{ir} = \left[\sum_{k=1}^{p} (x_{ik} - x_{rk})^2 \right]^{1/2}.$$

In termini vettoriali, la distanza euclidea può anche essere vista come la norma della differenza tra i 2 vettori riga della matrice dei dati. Infatti la (3.1) si può anche scrivere come:

$$d_{ir} = \left[(\mathbf{x}_i - \mathbf{x}_r)' (\mathbf{x}_i - \mathbf{x}_r) \right]^{1/2} = \left\| \mathbf{x}_i - \mathbf{x}_r \right\|.$$

Esempio 3.8
Si ipotizzi che per un gruppo di aziende siano stati rilevati gli indici di bilancio riportati nella seguente matrice dei dati:

Azienda	ROI	ROE	ROS	Indebitamento
A	7,1	8,2	5,7	25,3
B	0,3	1,5	2,8	53,4
C	0,4	2,5	5,2	34,1
D	5,1	2,0	1,2	11,7

La distanza euclidea tra la prima e la seconda azienda risulterà pari a:

$$d_{AB} = \left[(7{,}1-0{,}3)^2 + (8{,}2-1{,}5)^2 + (5{,}7-2{,}8)^2 + (25{,}3-53{,}4)^2\right]^{1/2} = 29{,}82$$

Procedendo in modo analogo per le altre coppie di righe, si ottiene la seguente matrice delle distanze euclidee:

$$D = \begin{bmatrix} 0 & 29{,}82 & 12{,}45 & 15{,}74 \\ 29{,}82 & 0 & 19{,}47 & 42{,}01 \\ 12{,}45 & 19{,}47 & 0 & 23{,}24 \\ 15{,}74 & 42{,}01 & 23{,}24 & 0 \end{bmatrix}.$$

La distanza euclidea presenta tuttavia due ordini di problemi. Anzitutto un problema di scala. Come si vede sia dalla formula che dall'esempio, vengono infatti sommate differenze al quadrato relative a caratteri diversi e misurati in unità di misura diverse. L'ipotesi che si adotta è che, ai fini della distanza tra le due unità, una differenza di una unità espressa nell'unità di misura di un carattere abbia la stessa importanza di una differenza di una unità espressa nell'unità di misura di un altro carattere. Riprendendo l'esempio precedente, la distanza tra due aziende viene calcolata assegnando alla differenza di un punto di ROI la stessa importanza della differenza di un punto nel grado di indebitamento, il che non appare appropriato. Ma altri esempi possono chiarire meglio come tale assunzione rappresenti una evidente forzatura. Si consideri, ad esempio, di aver rilevato in un gruppo di potenziali clienti il numero di figli e la cilindrata dell'automobile, misurata in centimetri cubi. In questo caso adottare la misura di distanza euclidea per confrontare due individui significa ipotizzare che una differenza di un figlio in più o in meno abbia la stessa importanza di una differenza di un centimetro cubo nella cilindrata dell'auto posseduta.
Per ovviare a tale problema si può ricorrere alla **standardizzazione** della matrice dei dati, depurando le variabili dall'effetto delle diverse unità di misura adottate e poi calcolare la distanza euclidea sui profili standardizzati.
La forma più comune di standardizzazione è quella che consiste nel sottrarre a ciascun elemento della matrice dei dati la media di colonna e dividere per la relativa deviazione standard, come nella espressione seguente:

$$z_{ik} = \frac{x_{ik} - \bar{x}_k}{\sigma_k}$$

La matrice dei dati standardizzati è di conseguenza adimensionale con tutti i vettori colonna che presentano media pari a zero e varianza unitaria.

Esempio 3.9
Riprendendo l'esempio precedente, si tratta di calcolare preliminarmente media aritmetica e deviazione standard per le quattro variabili rilevate:

Variabile	ROI	ROE	ROS	INDEB
Media	3,23	3,55	3,73	31,13
Deviazione standard	2,96	2,71	1,82	15,14

Il valore standardizzato corrispondente a x_{11} sarà quindi:

$$z_{11} = \frac{7{,}1 - 3{,}23}{2{,}96} = 1{,}31 .$$

In modo analogo possono essere calcolati i restanti valori ottenendo la seguente matrice Z dei dati standardizzati:

$$Z = \begin{bmatrix} 1{,}31 & 1{,}72 & 1{,}08 & -0{,}38 \\ -0{,}99 & -0{,}76 & -0{,}51 & 1{,}47 \\ -0{,}95 & -0{,}39 & 0{,}81 & 0{,}20 \\ 0{,}63 & -0{,}57 & -1{,}38 & -1{,}28 \end{bmatrix}.$$

La corrispondente matrice delle distanza euclidee calcolata a partire dalla matrice Z è, infine, la seguente:

$$D = \begin{bmatrix} 0 & 4{,}17 & 3{,}16 & 3{,}55 \\ 4{,}17 & 0 & 1{,}87 & 3{,}32 \\ 3{,}16 & 1{,}87 & 0 & 3{,}09 \\ 3{,}55 & 3{,}32 & 3{,}09 & 0 \end{bmatrix}.$$

Un secondo limite della distanza euclidea è che essa non tiene conto delle eventuali correlazioni esistenti tra le diverse variabili della matrice dei dati. Se due variabili risultano fortemente correlate significa che sono espressione dello stesso fenomeno e che nella misura della distanza (euclidea) tra le due unità si tiene conto due volte dello stesso fattore. Una possibile soluzione a questo secondo problema consiste nel calcolare la **distanza euclidea ponderata**, data dalla seguente espressione:

$$d_{ir} = \left[\sum_{k=1}^{p} (x_{ik} - x_{rk})^2 w_k \right]^{1/2}$$

dove w_k è un coefficiente di ponderazione.
In termini matriciali la stessa espressione si può scrivere nel modo seguente:

$$d_{ir} = \left[(\mathbf{x}_i - \mathbf{x}_r)' \mathbf{W} (\mathbf{x}_i - \mathbf{x}_r) \right]^{1/2}$$

dove **W** è una matrice diagonale (di dimensioni p x p) contenente i coefficienti di ponderazioni delle p variabili. Tali coefficienti dovrebbero risultare tanto maggiori quanto più la k-esima variabile è incorrelata con le altre p-1 variabili presenti nella matrice dei dati, cioè quanto più il contributo informativo della k-esima variabile è 'originale' rispetto a quello presente nelle variabili restanti. E al contrario, i coefficienti di ponderazione dovrebbero essere tanto minori quanto più la

k-esima variabile risulta correlata con le altre, in quanto il suo contributo informativo risulta almeno parzialmente 'duplicato' rispetto a quello apportato dalle altre variabili.
Invece di una matrice diagonale si può utilizzare una matrice **W** simmetrica piena[3]. Un caso particolare di distanza euclidea ponderata tramite una matrice simmetrica piena è la **distanza di Mahalanobis**, che assume come matrice di ponderazione l'inversa della matrice di covarianza. La sua espressione è dunque la seguente:

$$d_{ir} = \left[(\mathbf{x}_i - \mathbf{x}_r)' \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_r) \right]^{1/2}.$$

La distanza di Mahalanobis costituisce dunque una misura di distanza calcolata al netto della correlazione esistente tra le variabili. Allo stesso tempo essa elimina anche l'effetto derivante dalle diverse scale di misura adottate per le variabili e può essere quindi calcolata direttamente sulle variabili rilevate[4].

Esempio 3.10
Si consideri la seguente matrice dei dati contenente le variabili reddito mensile, spesa per consumo mensile e importo medio dello scontrino rilevate presso alcuni clienti di un esercizio commerciale:

reddito	consumo	Importo medio
2	1,5	10
4	3	9
3	1,8	11
1	1,3	9,5

A partire dalla matrice dei dati si calcolano le matrici *D* delle distanze euclidee, *S* delle covarianze e *R* delle correlazioni:

$$D = \begin{bmatrix} 0 & 2{,}69 & 1{,}45 & 1{,}14 \\ 2{,}69 & 0 & 2{,}54 & 3{,}48 \\ 1{,}45 & 2{,}54 & 0 & 2{,}55 \\ 1{,}14 & 3{,}48 & 2{,}55 & 0 \end{bmatrix};$$

$$S = \begin{bmatrix} 0{,}69 & 0{,}53 & -2{,}06 \\ 0{,}53 & 0{,}44 & -1{,}90 \\ -2{,}06 & -1{,}90 & 10{,}69 \end{bmatrix};$$

$$R = \begin{bmatrix} 1{,}00 & 0{,}92 & -0{,}08 \\ 0{,}92 & 1{,}00 & -0{,}46 \\ -0{,}08 & -0{,}46 & 1{,}00 \end{bmatrix}.$$

[3] In ogni caso, affinché si abbia $d_{ij} \geq 0$, la matrice W deve essere semi definita positiva
[4] Si può dimostrare infatti che la distanza di Mahalanobis è uguale anche alla distanza euclidea ponderata calcolata sulle variabili standardizzate utilizzando come matrice di ponderazione l'inversa della matrice di correlazione:

$$d_{ij} = \left[(\mathbf{x}_i - \mathbf{x}_j)' \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j) \right]^{1/2} = \left[(\mathbf{z}_i - \mathbf{z}_j)' \mathbf{R}^{-1} (\mathbf{z}_i - \mathbf{z}_j) \right]^{1/2}$$

Come era lecito attendersi, si rileva una forte correlazione positiva tra reddito e consumo, il che significa che le due variabili evidenziano un contenuto informativo almeno in parte comune. Supponiamo ora di voler calcolare la distanza euclidea tra la seconda e la quarta unità, quelle che presentano la massima differenza per il reddito e il consumo. Sommando differenze al quadrato tra 4 e 1 e poi tra 3 e 1,3 si sommano differenze che, almeno in parte, esprimono lo stesso fenomeno (la differenza di livello di benessere economico). La distanza euclidea tra la seconda e la quarta unità risulta infatti pari a 3,48, la distanza più grande rispetto a tutte le altre coppie di unità.
Calcoliamo ora la distanza di Mahalanobis tra le medesime due unità. Si deve innanzi tutto calcolare l'inversa della matrice delle covarianze, che è la seguente:

$$S^{-1} = \begin{bmatrix} 31{,}63 & -51{,}51 & -3{,}05 \\ -51{,}51 & 94{,}17 & 6{,}80 \\ -3{,}05 & 6{,}80 & 0{,}71 \end{bmatrix}.$$

Quindi la distanza di Mahalanobis tra le unità 2 e 4 sarà:

$$d_{24} = \left[(\mathbf{x}_2 - \mathbf{x}_4)' \mathbf{S}^{-1} (\mathbf{x}_2 - \mathbf{x}_4) \right]^{1/2} = \left[\begin{pmatrix} 3 & 1{,}7 & -0{,}5 \end{pmatrix} \begin{pmatrix} 31{,}63 & -51{,}51 & -3{,}05 \\ -51{,}51 & 94{,}17 & 6{,}80 \\ -3{,}05 & 6{,}80 & 0{,}71 \end{pmatrix} \begin{pmatrix} 3 \\ 1{,}7 \\ -0{,}5 \end{pmatrix} \right]^{1/2} = 5{,}40 \ .$$

Calcolando in modo analogo gli indici di distanza tra tutte le altre coppie di unità di ottiene la seguente matrice delle distanze di Mahalanobis:

$$D = \begin{bmatrix} 0 & 4{,}68 & 2{,}81 & 3{,}64 \\ 4{,}68 & 0 & 5{,}10 & 5{,}40 \\ 2{,}81 & 5{,}10 & 0 & 6{,}37 \\ 3{,}64 & 5{,}40 & 6{,}37 & 0 \end{bmatrix}.$$

Si può osservare come, utilizzando l'indice di Mahalanobis, la distanza tra la seconda e la quarta unità non è più la più elevata rispetto a tutte le altre coppie di unità: la massima distanza ora è quella tra la terza e la quarta unità.

Un indice di distanza più generale per caratteri quantitativi è dato dalla distanza di Minkowski, definito dalla seguente espressione:

Si osserva facilmente che quando $\lambda=2$ la distanza di Minkowski coincide con la distanza euclidea. Quando $\lambda=1$ la distanza di Minkowski viene detta distanza di Manhattan o distanza della città a blocchi, e assume la seguente espressione:

La definizione di distanza di Manhattan deriva dal fatto che in un esempio a due dimensioni (x1 e x2) la distanza tra due vertici di un edificio

Caratteri misti. Nell'analisi sui profili di colonna si è visto come non esista un unico indicatore che consenta di calcolare l'intensità e la direzione delle relazioni di associazione tra coppie di variabili, a prescindere dalla tipologia delle medesime (qualitative o quantitative) e utilizzando tutto il contenuto informativo presente nella matrice dei dati.
Nel caso delle relazioni tra unità una misura sintetica della distanza esistente tra due profili di riga di una matrice di dati con variabili di tipo misto si può invece ottenere. Tale misura è data dall'**indice di distanza di Gower**, definito nel modo seguente:

$$d_{ir} = \frac{\sum_{k=1}^{p} d_{ir,k}}{\sum_{k=1}^{p} \delta_{ir,k}},$$

dove $d_{ir,k}$ è una misura di distanza tra le righe i e r della matrice dei dati in relazione al k-esimo attributo, mentre $\delta_{ij,k}$ è una variabile dicotomica che assume valore uno se le due unità possono essere confrontate in relazione all'attributo k e zero altrimenti. La misura della distanza varia a seconda del tipo di carattere.

Caratteri quantitativi:

$$d_{ir,k} = \frac{|x_{ik} - x_{rk}|}{Range(k)}, \qquad \delta_{ir,k} = 1$$

dove *Range(k)* è il campo di variazione della variabile k.

Caratteri qualitativi ordinali:
Si trasformano le variabili in quantitative attribuendo punteggi crescenti al crescere delle modalità del carattere, riconducendoci al caso precedente.

Caratteri qualitativi dicotomici:
$d_{ir,k}$ assume valori 0 o 1 a seconda che le due unità presentino modalità uguali o diverse;
$\delta_{ir,k}$ assume sempre valore 1 salvo il caso di risposta negativa per entrambe le unità, nel qual caso assume valore 0.

Lo schema seguente riassume i valori assunti da $d_{ir,k}$ e $\delta_{ir,k}$ nei diversi casi:

$d_{ir,k}$	Unità i	
Unità r	Si	No
Si	0	1
No	1	0
$\delta_{ir,k}$	Unità i	
Unità r	Si	No
Si	1	1
No	1	0

Il fatto di porre $\delta_{ir,k}$ uguale a zero nel caso di assenza del fenomeno in entrambi i caratteri equivale ad adottare per questo tipo di caratteri la misura di distanza proposta da Jaccard.

Caratteri qualitativi sconnessi politomici:
Si assume come misura di distanza quella di Sneath; pertanto $d_{ir,k}$ assume valore uno se le unità i e j presentano modalità diversa in relazione al carattere k e zero altrimenti, mentre $\delta_{ir,k} = 1$.

La misura di distanza proposta da Gower è compresa tra zero e uno. Anche ad essa corrisponde dunque un indice di similarità definito come il complemento ad uno della distanza:

$c_{ir} = 1 - d_{ir}$

Esempio 3.11
Si supponga che su quattro clienti di una compagnia aerea siano stati rilevati il numero di voli nell'ultimo anno, il grado di soddisfazione, l'eventuale uso di internet per la prenotazione e il paese di residenza, ottenendo la seguente matrice dei dati:

Clienti	N. voli ultimo anno	Soddisfazione	Uso internet	Paese residenza
A	8	moltissimo	No	Italia
B	1	abbastanza	No	Francia
C	3	molto	Si	Francia
D	2	poco	Si	Germania

Per la soddisfazione si pone 1 = per niente, 2 = poco, 3 = abbastanza, 4 = molto e 5 = moltissimo.
Per l'uso di internet per la prenotazione, il fatto di aver risposto entrambi no, rende non confrontabili le unità, in quanto non possiamo sapere se la prenotazione è avvenuta con la medesima modalità o con modalità diversa, per cui $\delta_{ir,k}$ viene posto pari a zero.
La distanza tra le unità A e B è:

$$d_{AB} = \frac{\left[\frac{|8-1|}{8-1} + \frac{|5-3|}{5-1} + 0 + 1\right]}{1+1+0+1} = 0{,}83 \ ;$$

mentre quella tra A e C è:

$$d_{AC} = \frac{\left[\frac{|8-3|}{8-1} + \frac{|5-4|}{5-1} + 1 + 1\right]}{1+1+1+1} = 0{,}74 .$$

Procedendo analogamente per le altre coppie di variabili si ottiene la seguente matrice delle distanze di Gower:

$$D = \begin{bmatrix} 0 & 0{,}83 & 0{,}74 & 0{,}90 \\ 0{,}83 & 0 & 0{,}38 & 0{,}60 \\ 0{,}74 & 0{,}38 & 0 & 0{,}41 \\ 0{,}90 & 0{,}60 & 0{,}41 & 0 \end{bmatrix}.$$

3.6 Le analisi sui profili di colonna e di riga in Sas, R ed Xlstat

In SAS l'analisi sui profili di colonna può essere realizzata attraverso la procedura CORR. La procedura consente di calcolare la matrice di covarianza e la matrice di correlazione mediante l'indice di Bravais o il coefficiente di correlazione tra ranghi di Spearman. La sintassi fondamentale è la seguente:

PROC CORR DATA=nome	Specifica il sds di input; qualora l'opzione sia omessa viene Considerato l'ultimo data set creato nella sessione di lavoro.
PEARSON	Calcola la matrice di correlazione utilizzando il coefficiente di

	correlazione lineare di Bravais e Pearson.
SPEARMAN	Calcola la matrice di correlazione utilizzando il coefficiente di correlazione tra ranghi di Spearman; qualora le variabili presenti nel data set di input siano quantitative, esse vengono preventivamente trasformate in ranghi.
COV	Calcola la matrice di covarianze.
OUTP=nome	Specifica il sds di output dove viene memorizzata la matrice di correlazione (Bravais e Pearson).
OUTS=nome	Specifica il sds di output dove viene memorizzata la matrice di correlazione (Spearman).
VAR elenco variabili	Specifica l'elenco delle variabili sulle quali calcolare la matrice di correlazione; se omesso vengono considerate tutte le variabili numeriche del sas data set.
BY elenco variabili	Calcola tante matrici di correlazione sulla base dei gruppi formati dalle diverse modalità delle variabili che seguono lo statement BY.

In ambiente R per il calcolo della matrice di correlazione si può utilizzare la funzione cor; la funzione, seguita come argomento solo dalla matrice dei dati, calcola per default la matrice di correlazione utilizzando il coefficiente di correlazione lineare di Bravais e Pearson. Qualora si desideri calcolare il coefficiente di correlazione tra ranghi di Spearman occorre inserire l'opzione method="spearman". Per il calcolo dell'indice di associazione gamma di Goodman e Kruskal, in presenza di variabili ordinali con un ridotto numero di modalità, si può ricorrere alla funzione rcorr.cens, presente nel pacchetto Hmisc; in questo caso gli argomenti da inserire sono le due variabili per le quali si richiede il calcolo dell'indice gamma e l'opzione outx=TRUE. Nell'output l'indice gamma viene identificato con Dxy. Ovviamente, per riuscire ad acquisire l'ordinamento delle modalità dei caratteri qualitativi ordinali, le variabili dovranno essere definite come quantitative con modalità crescenti seguendo l'ordinamento delle variabili qualitative.

In Xlstat, una volta sistemata la matrice dei dati nel foglio Excel, per ottenere la matrice di correlazione si dovrà scegliere l'opzione *descrizione dei dati* (seconda opzione nel menù principale di Xlstat) e successivamente l'opzione *matrici di similarità/dissimilarità (correlazione)* (quarta opzione). Nel dialog box che si apre dovremo indicare la zona del foglio excel in cui è contenuta la matrice dei dati, scegliere l'opzione quantitativi per il tipo dei dati, indicare similarità e coefficiente di correlazione di Pearson per la scelta del tipo di prossimità, e di calcolare le prossimità per le colonne.

Esempio 3.12 – Procedure in Sas, R e Xlstat per il calcolo della matrice di correlazione

Una azienda di credito dispone delle informazioni sulla clientela aziendale relative al numero di addetti, alla durata del rapporto (anni) e al saldo (milioni di euro). I programmi seguenti, corredati dai relativi output, calcolano la matrice di correlazione utilizzando l'indice di Bravais e Pearson in Sas, R e Xlstat.

```
* esempio procedura CORR;

data banca;
   input n_addetti durata saldo;
       cards;
         8   4   0.2
         5   1   0.8
         12  9  -0.9
         82  4  -1.5
         3   1  -0.3
         43  5   1.3
```

```
          12   7   -0.2
          21   8   -0.6
          54   3   -1.7
          4    2   -1.4
;
proc corr data=banca pearson outp=risu ;
run;
```

La procedura CORR

3 Variabili: n_addetti durata saldo

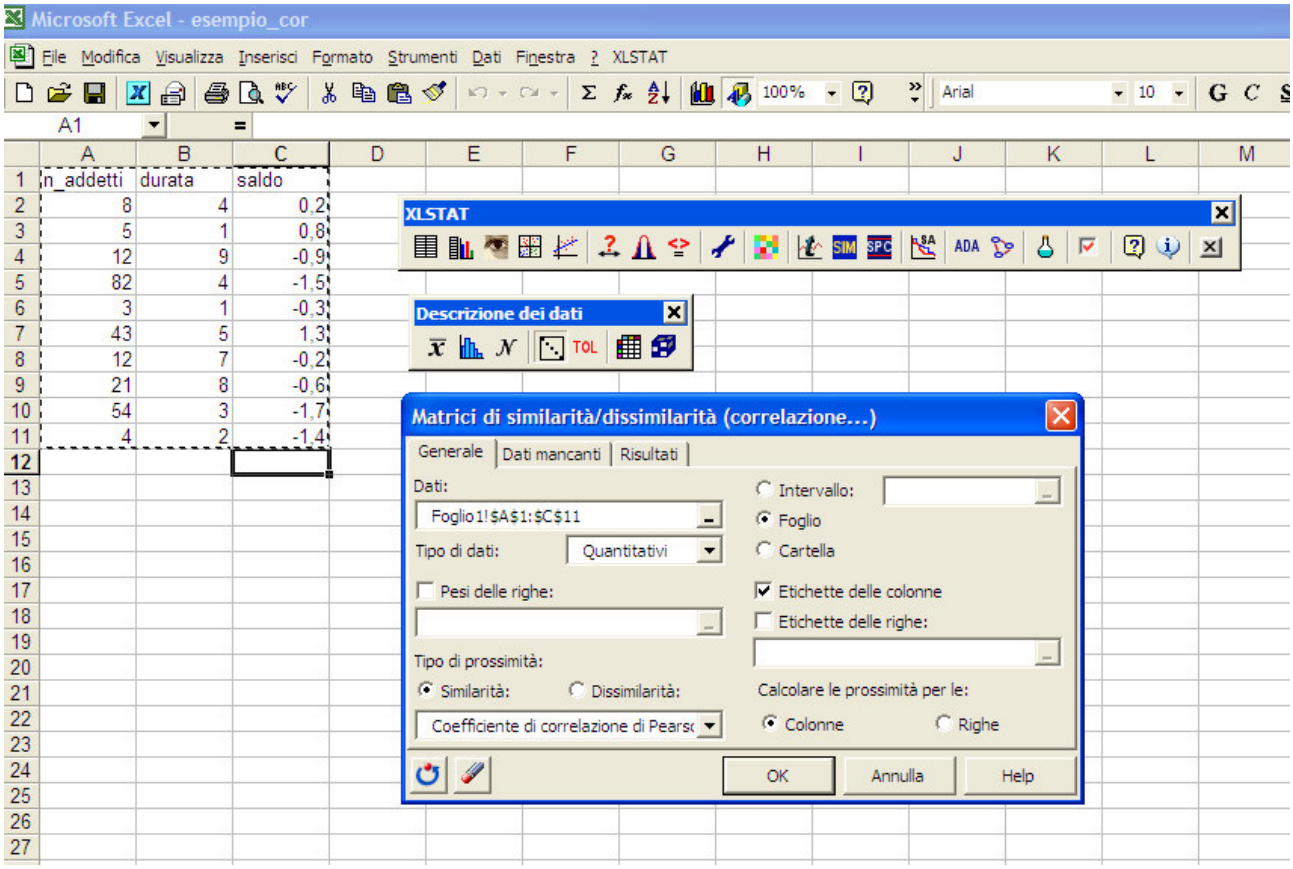
Statistiche semplici

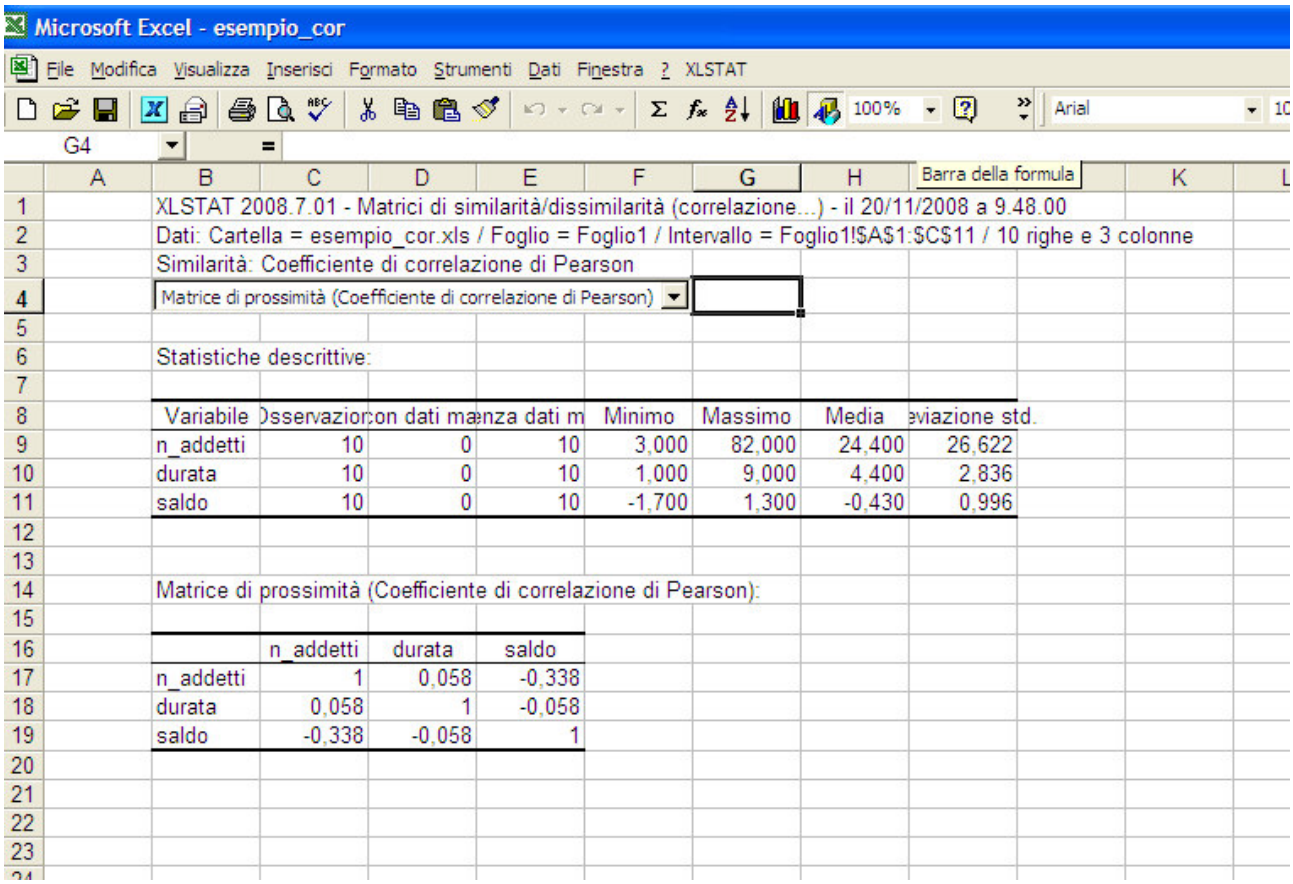
Variabile	N	Media	Dev std	Somma	Minimo	Massimo
n_addetti	10	24.40000	26.62163	244.00000	3.00000	82.00000
durata	10	4.40000	2.83627	44.00000	1.00000	9.00000
saldo	10	-0.43000	0.99560	-4.30000	-1.70000	1.30000

Coefficienti di correlazione di Pearson, N = 10
Prob > |r| con H0: Ro=0

	n_addetti	durata	saldo
n_addetti	1.00000	0.05798 0.8736	-0.33822 0.3391
durata	0.05798 0.8736	1.00000	-0.05824 0.8730
saldo	-0.33822 0.3391	-0.05824 0.8730	1.00000

```
> dati
   n_addetti durata saldo
1          8      4   0.2
2          5      1   0.8
3         12      9  -0.9
4         82      4  -1.5
5          3      1  -0.3
6         43      5   1.3
7         12      7  -0.2
8         21      8  -0.6
9         54      3  -1.7
10         4      2  -1.4
> cor(dati)
            n_addetti      durata       saldo
n_addetti  1.00000000  0.05797905 -0.33822308
durata     0.05797905  1.00000000 -0.05823521
saldo     -0.33822308 -0.05823521  1.00000000
>
```


Microsoft Excel - esempio_cor
File Modifica Visualizza Inserisci Formato Strumenti Dati Finestra ? XLSTAT
A1
n_addetti durata saldo
8 4 0,2
5 1 0,8
12 9 -0,9
82 4 -1,5
3 1 -0,3
43 5 1,3
12 7 -0,2
21 8 -0,6
54 3 -1,7
4 2 -1,4
XLSTAT
Descrizione dei dati
Matrici di similarità/dissimilarità (correlazione...)
Generale
Dati mancanti
Risultati
Dati:
Foglio1!$A$1:$C$11
Tipo di dati:
Quantitativi
Pesi delle righe:
Tipo di prossimità:
Similarità:
Dissimilarità:
Coefficiente di correlazione di Pears
Intervallo:
Foglio
Cartella
Etichette delle colonne
Etichette delle righe:
Calcolare le prossimità per le:
Colonne
Righe
OK
Annulla
Help


Microsoft Excel - esempio_cor
File Modifica Visualizza Inserisci Formato Strumenti Dati Finestra ? XLSTAT
G4
Barra della formula
XLSTAT 2008.7.01 - Matrici di similarità/dissimilarità (correlazione...) - il 20/11/2008 a 9.48.00
Dati: Cartella = esempio_cor.xls / Foglio = Foglio1 / Intervallo = Foglio1!$A$1:$C$11 / 10 righe e 3 colonne
Similarità: Coefficiente di correlazione di Pearson
Matrice di prossimità (Coefficiente di correlazione di Pearson)
Statistiche descrittive:
Variabile Osservazioni con dati mancanza dati m Minimo Massimo Media eviazione std.
n_addetti 10 0 10 3,000 82,000 24,400 26,622
durata 10 0 10 1,000 9,000 4,400 2,836
saldo 10 0 10 -1,700 1,300 -0,430 0,996
Matrice di prossimità (Coefficiente di correlazione di Pearson):
n_addetti durata saldo
n_addetti 1 0,058 -0,338
durata 0,058 1 -0,058
saldo -0,338 -0,058 1

Per quanto attiene alle analisi sui profili di riga, in Package Sas mette a disposizione la procedura DISTANCE, la cui sintassi essenziale è la seguente:

PROC DISTANCE	DATA=nome	Specifica il sds di input; qualora l'opzione sia omessa viene Considerato l'ultimo data set creato nella sessione di lavoro.
	OUT=nome	Specifica il sas data set di output in cui viene memorizzata la matrice di distanza.
	METHOD=	Specifica il metodo di calcolo della matrice di distanza; le opzioni possibili sono:
		Euclid — distanza euclidea
		L(p) — distanza di Minkowski con λ=p
		Cityblock — L(1) o distanza di Manhattan
		Chebychev — L(∞) o distanza di Lagrange
		Dsqmatch — Indice di Sneath o Simple matching se dati dicotomici
		Djaccard — Indice di Jaccard
		Dgower — Indice di Gower per caratteri misti.
VAR		Specifica l'elenco e il tipo delle variabili sulle quali calcolare La matrice di distanza.La sintassi è la seguente:

VAR livello (elenco variabili / opzioni).
I livelli sono: ANOMINAL, ORDINAL, INTERVAL o RATIO; le opzioni sono STD=metodo di standardizzazione e WEIGHT= pesi delle variabili. Per quanto attiene ai metodi di standardizzazione il più utilizzato e richiamato anche nel paragrafo 3.5 è STD.

BY elenco variabili	Calcola tante matrici di distanza sulla base dei gruppi formati dalle diverse modalità delle variabili che seguono lo statement BY.

In ambiente R per il calcolo della matrice di distanza si può utilizzare la funzione dist appartenente alla libreria mva. La sintassi essenziale della funzione dist è la seguente:
dist (dati , method="euclidean" , diag = FALSE , upper = FALSE)
dove:
dati è una matrice o un data frame contenente la matrice dei dati;
method specifica l'indice da utilizzare per il calcolo della distanza: le opzioni disponibili sono:

- euclidean, per il calcolo della distanza euclidea;
- manhattan, per il calcolo dell'indice di distanza di Manhattan o della città a blocchi;
- maximum, per il calcolo della distanza di Lagrange;
- minkowski, per il calcolo della distanza di Minkowski, in questo caso occorre anche l'opzione p=λ;
- canberra, per il calcolo della distanza di Canberra;
- binary, per il calcolo della distanza di Jaccard, in presenza di variabili dicotomiche;

le opzioni diag e upper vengono poste uguali a TRUE quando si richiede che la matrice di distanza contenga rispettivamente i valori pari a zero sulla diagonale e la parte triangolare superiore della matrice; in assenza di tali opzioni il risultato e una matrice triangolare inferiore senza la diagonale principale.
Per il calcolo della matrice di distanza in presenza di caratteri misti, basata sull'indice di Gower, è disponibile la funzione daisy, appartenente alla libreria cluster, caratterizzata dalla seguente sintassi:
daisy(dati, metric = c("gower"), type = list())
dove:

dati è una matrice o un data frame contenente la matrice dei dati. Qualora le variabili siano alcune qualitative e altre quantitative la funzione utilizza per default l'indice di Gower senza bisogno dell'opzione metric;
metric specifica l'indice da utilizzare per il calcolo della distanza: oltre a Gower, sono possibili anche le opzioni euclidean e manhattan.
Per il calcolo della distanza di Mahalanobis è disponibile la funzione mahalanobis, sempre nella libreria cluster.

In Xlstat, una volta sistemata la matrice dei dati nel foglio Excel, per ottenere la matrice di correlazione si dovrà scegliere l'opzione *descrizione dei dati* (seconda opzione nel menù principale di Xlstat) e successivamente l'opzione *matrici di similarità/dissimilarità (correlazione)* (quarta opzione). Nel dialog box che si apre dovremo indicare la zona del foglio excel in cui è contenuta la matrice dei dati, scegliere l'opzione quantitativi per il tipo dei dati, indicare similarità e coefficiente di correlazione di Pearson per la scelta del tipo di prossimità, e di calcolare le prossimità per le colonne.

Esempio 3.13 – Procedure in Sas, R e Xlstat per il calcolo della matrice di distanza

Un esercizio commerciale della grande distribuzione dispone delle informazioni sui possessori di carta fedeltà relative alla professione, alla dimensione della famiglia, alla spesa media mensile e al sesso. I programmi seguenti, corredati dai relativi output, calcolano la matrice di correlazione utilizzando l'indice di Bravais e Pearson in Sas, R e Xlstat.

```
data carta_fede;
   input professione $ n_compo spesa sesso $;
       cards;
Impiegato   2  250 M
Operaio     5  300 F
Operaio     1  250 M
Operaio     3  350 M
Lav.aut.    1  320 F
Impiegato   2  400 F
Operaio     3  390 M
Lav.aut.    2  400 M
Impiegato   2  390 F
;
proc distance data=carta_fede method=dgower out=dista;
     var anominal(professione sesso) ratio(n_compo spesa);
run;
proc print data=dista;
run;
```

Oss	Dist1	Dist2	Dist3	Dist4	Dist5	Dist6	Dist7	Dist8	Dist9
1	0.00000	.	.	.	.	.	.	.	.
2	0.77083	0.00000	.	.	.	.	.	.	.
3	0.31250	0.58333	0.00000	.	.	.	.	.	.
4	0.47917	0.45833	0.29167	0.00000	.	.	.	.	.
5	0.67917	0.53333	0.61667	0.67500	0.00000	.	.	.	.
6	0.50000	0.60417	0.81250	0.64583	0.44583	0.00000	.	.	.

```
    7    0.54583   0.52500   0.35833   0.06667   0.74167   0.57917   0.00000
.          .
    8    0.50000   0.85417   0.56250   0.39583   0.44583   0.50000   0.32917
0.00000      .
    9     0.48333    0.58750    0.79583    0.62917    0.42917    0.01667    0.56250
0.51667      0
```

```
> dati
  professione n_compo spesa sesso
1   Impiegato   2        250    M
2   Operaio      5        300    F
3   Operaio      1        250    M
4   Operaio      3        350    M
5   Lav.aut.     1        320    F
6   Impiegato   2        400    F
7   Operaio      3        390    M
8   Lav.aut.     2        400    M
9   Impiegato    2        390    F
> daisy(dati)
Dissimilarities :
          1         2         3         4         5         6         7
2 0.77083333
3 0.31250000 0.58333333
4 0.47916667 0.45833333 0.29166667
5 0.67916667 0.53333333 0.61666667 0.67500000
6 0.50000000 0.60416667 0.81250000 0.64583333 0.44583333
7 0.54583333 0.52500000 0.35833333 0.06666667 0.74166667 0.57916667
8 0.50000000 0.85416667 0.56250000 0.39583333 0.44583333 0.50000000 0.32916667
9 0.48333333 0.58750000 0.79583333 0.62916667 0.42916667 0.01666667 0.56250000
          8
2
3
4
5
6
7
8
9 0.51666667
```

Bibliografia essenziale

Fabbris, L. (1997), *Statistica multivariata. Analisi esplorativa dei dati*, Mc-Graw Hill, Milano.

Fabbris, L. (1997), *Statistica multivariata. Analisi esplorativa dei dati*, Mc-Graw Hill, Milano.

Kruskal, J.B. (1964), *Non-metric multidimensional scaling: a numerical method*, in *Psycometrica*, 29, pp. 209-229.

Kruskal, J.B. (1965), *Analysis of factorial experiments by estimating monotone transformations of the data*, in *Journal of Royal Statistical Society*, 27, pp. 251-263.

Lancaster, K. (1966), *A new approach to consumer theory*, in *Journal of Political economics*, 74, pp. 132-157.