Capitolo I

LE MATRICI DEI DATI E LE RELAZIONI TRA LE VARIABILI

> «*Non ci sono informazioni migliori delle altre. Il potere sta nello schedarle tutte e poi cercare le connessioni. Le connessioni ci sono sempre, basta volerle trovare*»
>
> (Umberto Eco, *Il pendolo di Foucault*).

1. *La matrice dei dati «unità per variabili»*

Il punto di partenza abituale delle analisi statistiche è una tabella che per le n unità statistiche considerate (individui, oggetti, aziende, etc.) presenta le modalità di p variabili quantitative e/o qualitative (anche dette fenomeni o caratteri statistici). Nel caso di variabili quantitative le modalità sono espresse con numeri (che si ottengono da operazioni di misurazione o di conteggio), mentre nel caso di variabili qualitative le modalità sono solitamente espresse con parole. In quest'ultima circostanza può essere utile sostituire le modalità con opportuni codici.

Le informazioni rilevate vengono quindi presentate nella seguente matrice dei dati del tipo «unità per variabili», di dimensioni $n \times p$:

$$\mathbf{X} = [x_{is}]$$

ove: x_{is} = modalità quantitativa (ovvero codice) che nella unità statistica i-esima ($i = 1, 2, ..., n$) presenta la variabile s-esima ($s = 1, 2, ..., p$).

Esempio. Una matrice dei dati molto semplice è presentata nella tab. 1.1. (1). In essa, con riferimento ai forni a microonde di 8 marche

---

(1) In questo libro adotteremo il punto come separatore decimale, secondo

diverse (unità statistiche) sono riportati i valori di tre variabili quantitative (altezza interna, in cm; capacità, in litri; prezzo, in migliaia di lire) ed i codici di due fenomeni qualitativi, il primo di tipo nominale dicotomico (timer, con modalità: meccanico, codificato 1, oppure digitale, codificato 2) ed il secondo ordinale, che rappresenta i giudizi su una scala da 1 a 10 formulati da un esperto sulla qualità della cottura. (Fonte dei dati: *Altro consumo*, n. 104, aprile 1998).

TAB. 1.1. *Caratteristiche di 8 forni a microonde.*

	MARCA	ALTEZZA	CAPACITA'	PREZZO	TIMER	COTTURA
1	CANDY	20	15.7	295	1	9
2	DELONGHI	17	10.2	260	1	10
3	ELECTROLUX	19	12.5	259	2	7
4	MOULINEX	16	10.8	280	2	8
5	OCEAN	20	10.5	279	1	4
6	PANASONIC	19	11.5	240	2	6
7	SAMSUNG	17	11.8	259	2	6
8	SHARP	17	10.8	339	2	7

Sotto l'aspetto informatico la matrice dei dati rappresenta il *file* iniziale sul quale si effettueranno le successive elaborazioni. Utilizzando il *package* statistico SPSS (*Statistical Package for the Social Sciences*) — al quale faremo sempre riferimento in seguito, nella versione 8.0 — i dati contenuti nel *file* possono essere immessi direttamente dalla tastiera oppure possono essere «importati» come *file* scritto con un diverso *software* (ad esempio, Excel, Lotus, dBASE) (2).

Nel testo: S. Zani, *Analisi dei dati statistici*, vol. I, *Osservazioni in una e due dimensioni*, Giuffrè, Milano, 1997 (nel seguito richiamato semplicemente come Vol. I) sulla matrice dei dati abbiamo condotto le analisi di tipo unidimensionale, cioè riferite ad una singola variabile (calcolo di medie, indici di variabilità, indici della forma di distribuzione) e bidimensionale (correlazione, cograduazione ed associazione).

---

l'impostazione internazionale, poiché questa è la convenzione prevalente nei pacchetti (*packages*) statistici.

(2) Per un'introduzione a SPSS rinviamo, oltre che ai relativi manuali, a Gerber and Voelkl (1997); Voelkl and Gerber (1999).

In questo volume esamineremo le analisi multidimensionali, cioè i metodi per lo studio simultaneo d'una pluralità di variabili, secondo un approccio prevalentemente esplorativo (Vol. I, pp. 7-10) (3).

Per chiarire la distinzione fondamentale tra le due categorie di dati che si possono analizzare, consideriamo due esempi di *data sets* multidimensionali:

1) gli n clienti d'una azienda, per ciascuno dei quali si conoscono le modalità di p variabili quantitative (numero di ordini effettuati in un anno, ammontare degli acquisti in Euro, etc.) e qualitativi (regione di residenza, tipologia del cliente, etc.);

2) gli n individui che costituiscono un campione probabilistico estratto da un certo universo e per i quali si conoscono le risposte (opportunamente codificate) fornite a p domande (quantitative e/o qualitative) d'un questionario.

I due esempi corrispondono alle situazioni che si possono presentare nelle analisi statistiche: nel primo esempio i dati rilevati costituiscono l'intera *popolazione* d'interesse, per cui il loro studio può esaurire le finalità della ricerca; nel secondo caso, invece, l'esame delle caratteristiche delle unità del *campione* è la premessa di un procedimento inferenziale, che aspira a raggiungere conoscenze, in termini probabilistici, sull'universo statistico corrispondente.

L'obiettivo generale dell'analisi dei dati è la ricerca di configurazioni o strutture (*patterns*) esistenti nelle informazioni rilevate, e si può articolare nei seguenti punti:

1) individuazione di valori anomali (*outliers*) multivariati, che possono distorcere l'interpretazione dei dati ed inficiare le fasi successive dell'analisi;

2) ricerca delle relazioni, anche non lineari, esistenti tra le variabili, come premessa alla formulazione e all'adattamento di modelli;

3) riduzione delle dimensioni dello spazio in cui figurano le unità, R^p, al fine di fornire una sintesi delle relazioni tra le variabili e di consentirne anche opportune rappresentazioni grafiche nel piano o in R^3;

---

(3) Citiamo alcuni testi in lingua italiana che trattano gli argomenti di questo libro: Fraire (1994); Mignani e Montanari (1994); Rizzi (1995); Fabbris (1997). Per un'ampia panoramica dei contributi più recenti, con particolare riguardo alla scuola italiana, si vedano le rassegne di Balbi (1994); Balbi e Lauro (1997).

4) misura della diversità, ovvero della rassomiglianza, tra coppie di unità statistiche, con riferimento ai valori e/o alle modalità qualitative che presentano le p variabili considerate;

5) classificazione delle unità in gruppi omogenei (ottenuti in base alle precedenti misure di diversità e rassomiglianza o con altri criteri), nell'intento di ottenere un numero ridotto di categorie o classi.

Nel presente volume, pur illustrando separatamente i diversi metodi, cercheremo di porre in luce che l'analisi dei dati è un processo iterativo ed integrato. Infatti, nell'insieme delle informazioni di partenza — che può essere molto vasto, eterogeneo e con dati di differente qualità — si cerca in primo luogo d'individuare con le analisi multidimensionali il nucleo (*core*) di quelle di reale interesse nell'indagine che si sta conducendo. Questo può comportare l'eliminazione di unità anomale e di variabili ridondanti o inutili. Sul sottoinsieme ottenuto si riapplicano i metodi multivariati ed i risultati dei medesimi possono suggerire ulteriori aggiustamenti del *data set*. Inoltre, sul medesimo insieme di dati si applicano solitamente diverse metodologie, per mettere in evidenza tutti i possibili aspetti del *pattern* sottostante. La concordanza tra i risultati ottenuti con metodi differenti costituisce un importante elemento di supporto alla validità delle conclusioni raggiunte (4).

In molte circostanze, infine, le conoscenze ricavate dall'analisi dei dati sono una premessa per la formulazione e la costruzione di modelli probabilistici, che ambiscono a fornire uno schema interpretativo di validità generale dei fenomeni osservati.

---

(4) In tempi molto recenti è stata coniata una nuova denominazione per le analisi statistiche (esplorative) riguardanti insiemi molto numerosi: *data mining*. L'aspetto di novità è costituito dal riferimento a basi di dati assai vaste, generalmente ottenute attraverso processi automatici o elettronici di raccolta delle informazioni. Si pensi, ad esempio, ad una catena di supermercati, ove la vendita di ogni articolo è registrata attraverso un sistema EPOS (*Electronic Point of Sale*), oppure ad una compagnia di telecomunicazioni, che memorizza gli scatti telefonici di ciascuno dei propri utenti, o ancora ad una banca che rileva tutte le operazioni effettuate dai clienti presso tutte le proprie filiali. Con il termine *data mining* si indica il processo di identificazione in un *database* di strutture dei dati che siano valide ed utili dal punto di vista operativo (Fayyad *et al.*, 1996). Esso può anche essere definito come «*heterogeneous collection of tools for extracting the potentially valuable information in data mountains*» (Hand, 1998). Un sinonimo di *data mining*, utilizzato da alcuni Autori, è *knowledge discovery in data bases*.

2. *La matrice degli scostamenti standardizzati*

Con riferimento alle sole variabili quantitative (come si è visto nel Vol. I, pp. 147-149), la matrice dei dati può essere trasformata nella matrice $\mathbf{X}$ degli scostamenti (o scarti) dalla media, talvolta chiamata matrice dei dati « centrata », in cui le medie di ciascuna delle colonne sono uguali a zero. Inoltre, la matrice dei dati può essere trasformata nella matrice $\mathbf{Z}$ espressa in termini di scostamenti standardizzati:

$$\mathbf{Z} = [z_{is}]$$

ove: z_{is} = scostamento standardizzato per l'unità statistica i-esima e la variabile s-esima.

Com'è noto, gli scostamenti standardizzati d'una variabile sono puri numeri, hanno media nulla e varianza unitaria e pertanto sono comparabili anche per variabili originariamente espresse in unità di misura differenti e/o con diverso ordine di grandezza.

TAB. 1.2. *Medie e scostamenti quadratici medi delle prime 3 variabili della tab. 1.1.*

	N	Media	Deviazione std.
ALTEZZA	8	18.13	1.55
CAPACITA'	8	11.73	1.77
PREZZO	8	276.38	30.43

Esempio. Per calcolare le medie e gli scostamenti quadratici medi (*standard deviations*) delle prime tre variabili (quantitative) della tab. 1.1 occorre scegliere nel menù di SPSS la sequenza: *statistica-riassumi-descrittive.* Si selezionano le variabili d'interesse e le eventuali opzioni (è possibile calcolare anche la varianza, il minimo ed il massimo, gli indici di asimmetria e curtosi). Inoltre, per determinare gli scostamenti standardizzati, si sceglie: *salva valori standardizzati come variabili.* Il programma aggiunge i valori standardizzati di ognuna delle variabili selezionate come ulteriori colonne nella matrice dei dati iniziale, denominandole automaticamente come z-(nome variabile). Con riferimento alla tab. 1.1, riportiamo le medie e le deviazioni standard per le tre variabili quantitative nella tab. 1.2 e gli scostamenti standardizzati nella tab. 1.3.

TAB. 1.3. *Scostamenti standardizzati delle prime 3 variabili della tab. 1.1.*

	MARCA	Z-ALTEZZA	Z-CAPACITA'	Z-PREZZO
1	CANDY	1.208	2.242	.612
2	DELONGHI	-.725	-.860	-.538
3	ELECTROLUX	.564	.437	-.571
4	MOULINEX	-1.369	-.522	.119
5	OCEAN	1.208	-.691	.086
6	PANASONIC	.564	-.127	-1.195
7	SAMSUNG	-.725	.042	-.571
8	SHARP	-.725	-.522	2.058

Ad esempio, l'altezza, del forno Candy risulta maggiore della media in misura pari a 1.208 volte lo scostamento quadratico medio di tale variabile, mentre il forno DeLonghi ha un'altezza minore della media di 0.725 volte la *standard deviation.*

3. *La ponderazione delle unità statistiche*

Se le unità statistiche hanno una diversa dimensione o un'importanza differente è molto naturale attribuire a ciascuna di esse un differente «peso». In tal caso, oltre alla matrice dei dati che contiene la variabili d'interesse, occorre dunque considerare anche il vettore dei pesi. Può essere utile esprimere i pesi in termini relativi (rapportando ciascun peso alla somma dei pesi), pervenendo al seguente vettore:

$$\mathbf{w} = [w_1, ..., w_i, ..., w_n]' \text{ con } \sum_{i=1}^{n} w_i = 1.$$

Per alcuni sviluppi algebrici utilizzati successivamente, tale vettore può essere trascritto nella forma della seguente matrice diagonale, di dimensioni $n \times n$:

$$\mathbf{W} = diag[w_i].$$

Esempio. Si consideri la tab. 1.4, riferita a 10 aziende del settore alimentare. Per ciascuna di esse si sono rilevati i valori dei seguenti indicatori economici (fonte: *Le 5000 società leader, Supplemento a Milano Finanza*, 1998, dati riferiti all'anno 1997):

— ECON.PRO = *economic profit*, or *loss* (differenziale tra il rendimento del capitale investito ed il suo costo, in miliardi di lire);
— CASH.FAT = *cash flow* sul fatturato, in %;
— LAVOR.VA = costo del lavoro sul valore aggiunto, in %;
— ROE = *return on equity* (utile netto sul patrimonio, in %);
— INDE.CAP = indebitamento sul capitale proprio.

TAB. 1.4. *Indicatori economici di 10 aziende alimentari.*

	AZIENDA	ECON.PRO	CASH.FAT	LAVOR.VA	ROE	INDE.CAP	FATTURATO
1	BARILLA	-25.40	7.39	59.54	4.20	.83	2867
2	ERIDANIA	-141.00	4.00	68.99	.84	1.86	1693
3	FERRERO	65.80	9.61	53.70	21.12	-.02	3031
4	GALBANI	-71.90	8.40	56.32	2.66	-.02	2136
5	KRAFT	-32.00	5.88	72.11	3.20	.35	1563
6	LAVAZZA	-28.90	4.96	39.08	5.29	-.05	1117
7	NESTLE'	-98.80	2.72	81.25	.00	1.69	3463
8	PARMALAT	-145.10	5.96	38.51	2.23	2.91	1664
9	PLASMON	31.70	27.76	31.35	24.60	1.35	858
10	STAR	2.40	6.47	62.49	10.60	.00	811

Tali aziende potrebbero essere pesate con criteri differenti: il numero di dipendenti, il fatturato, il valore aggiunto, etc. In relazione alle finalità dell'indagine si sceglie la variabile di ponderazione ritenuta più opportuna; nella tab. 1.4 si è adottato come peso il FATTURATO, in miliardi di lire.

TAB. 1.5. *Statistiche descrittive ponderate delle variabili della tab. 1.4.*

	N	Media	Deviazione std.
ECON.PRO	19203	-46.9935	68.2728
CASH.FAT	19203	7.1948	5.0082
LAVOR.VA	19203	59.8840	14.5157
ROE	19203	6.6388	8.0308
INDE.CAP	19203	.9253	.9445
Validi (listwise)	19203		

Per calcolare gli indici sintetici ponderati occorre scegliere dal menù: *pesa i casi*, indicando quindi la variabile di ponderazione (nell'esempio il FATTURATO). Si procede quindi come in precedenza con la sequenza: *statistica-riassumi-descrittive*, ottenendo la tabella delle statistiche descrittive (nella quale N indica ora la somma dei pesi utilizzati) (v. tab. 1.5).

4. *La matrice dei dati partizionata*

Nelle analisi multidimensionali può essere interessante considerare sottoinsiemi delle *n* unità statistiche. Si pensi, ad esempio, agli studenti della facoltà di Economia di Parma, suddivisi per anno di corso oppure alle aziende iscritte alla Camera di Commercio distinte per settore di appartenenza.

Se i sottoinsiemi sono a due a due disgiunti e la loro unione coincide con l'insieme di partenza, i gruppi di unità costituiscono una *partizione* dell'insieme originario (vol. I, pp. 41-44).

Con riferimento ad una partizione delle unità, precisiamo che essa può essere:

i) *nota a priori*, in base ad una caratteristica delle unità che risulta conosciuta prima di effettuare l'indagine statistica. Ad esempio, se le unità statistiche sono i comuni d'una regione, è ben nota, e molto naturale, la partizione degli stessi nelle province;

ii) *determinata in base ad uno o più caratteri che compaiono nella matrice dei dati rilevati.* Ad esempio, in un sondaggio d'opinione, solo dopo l'effettuazione dello stesso si potrà proporre la partizione dei rispondenti nei tre gruppi: favorevoli ad un certo provvedimento, contrari, «non so». Si potranno poi studiare, distintamente per i tre gruppi, le altre variabili rilevate;

iii) *ottenuta in base ad un metodo di* cluster analysis, *che riunisce in gruppi le unità tra loro simili* (si vedano i successivi cap. 5 e 6).

In ogni caso, tuttavia, la considerazione d'una certa partizione della matrice dei dati discende da una scelta del ricercatore, che adotta un proprio angolo visuale per l'analisi e vuole porre in luce taluni aspetti, trascurandone eventualmente molti altri. Ad esempio, per alcune finalità i comuni d'una regione potrebbero essere partizionati — anziché per provincia — per classi di dimensione demografica (sino a 1000 abitanti, da 1001 a 5000, oltre 5000); i rispondenti ad un sondaggio potrebbero essere suddivisi per sesso, oppure partizionati congiuntamente per sesso e per classe d'età, etc.

Esempio. Una matrice partizionata è riportata nella tab. 1.6, che per le regioni italiane presenta i valori dei seguenti indicatori inerenti alle forze di lavoro (fonte: Istat, *Forze di lavoro - Media 1997*, Roma, 1998):

— CERCA = % di persone in cerca di occupazione sulle forze di lavoro;
— FORZE = % delle forze di lavoro sulla popolazione residente;
— LAUREA = % di laureati sulle forze di lavoro;
— DIPLOMA = % di diplomati sulle forze di lavoro;
— TERZIARI = % di occupati nel terziario sulle forze di lavoro.

TAB. 1.6. *Indicatori inerenti alle forze di lavoro nelle regioni italiane.*

	REGIONE	ZONA	ABITANTI	CERCA	FORZE	LAUREA	DIPLOMA	TERZIARI
1	Piemonte	NC	4291441	8.3	43.7	8.2	28.5	55.0
2	Valle d'A.	NC	119610	5.6	46.6	7.3	29.1	69.2
3	Lombardia	NC	8988951	6.1	44.1	9.9	28.8	55.4
4	Trentino	NC	924281	3.9	45.8	6.1	22.8	63.0
5	Veneto	NC	4469156	5.6	44.4	7.2	25.8	53.2
6	Friuli V.G.	NC	1184654	7.1	42.7	9.0	29.3	60.4
7	Liguria	NC	1641835	11.7	39.9	11.1	31.4	73.8
8	Emilia-R.	NC	3947102	5.6	46.2	9.8	29.0	57.9
9	Toscana	NC	3527303	8.4	42.4	9.0	29.2	61.6
10	Umbria	NC	831714	10.2	40.0	9.1	35.5	62.5
11	Marche	NC	1450879	6.5	42.4	8.9	28.4	53.9
12	Lazio	NC	5242709	12.8	40.2	14.0	36.5	75.7
13	Abruzzo	M	1276040	9.6	39.0	10.2	30.9	59.3
14	Molise	M	329894	16.5	39.0	8.7	30.7	57.1
15	Campania	M	5796899	25.5	35.0	10.2	29.2	66.1
16	Puglia	M	4090068	17.7	34.5	8.9	27.1	61.9
17	Basilica	M	610330	18.1	35.2	7.5	26.8	54.1
18	Calabria	M	2070992	25.2	33.8	10.8	31.3	67.2
19	Sicilia	M	5108067	23.5	33.8	10.4	29.8	67.8
20	Sardegna	M	1661429	21.0	37.8	7.7	25.1	64.4

Seguendo il criterio adottato dall'Istat, le regioni sono suddivise nelle due grandi ripartizioni geografiche: Nord-Centro (NC) e Mezzogiorno (M) (5). È chiaro che un'analisi della situazione occupazionale

(5) Una partizione più «fine» delle regioni italiane è la seguente: Italia Nord-Occidentale (Piemonte, Valle d'Aosta, Lombardia, Liguria); Italia Nord-Orientale (Trentino-Alto Adige, Veneto, Friuli-Venezia Giulia, Emilia-Romagna); Italia Cen-

può essere condotta, oltre che sull'insieme delle regioni italiane, sulle regioni appartenenti a ciascuna delle aree suddette.

Per effettuare i calcoli distintamente per ogni gruppo di unità, in SPSS occorre scegliere dal menù: *distingui* e l'opzione: *distingui i risultati per gruppo*, indicando la variabile categorica in base alla quale si intendono formare i sottoinsiemi (nell'esempio in esame, la variabile ZONA). Inoltre, volendo effettuare l'analisi ponderata, si sceglie come in precedenza *pesa i casi*, adottando ora come pesi gli ABITANTI. Procedendo quindi con i comandi: *statistica-riassumi-descrittive*, si ottengono gli indici statistici unidimensionali ponderati, con riferimento a ciascuna delle due ripartizioni geografiche (v. tab. 1.7a, 1.7b).

TAB. 1.7a. *Statistiche descrittive delle variabili della tab. 1.6: regioni del Nord-Centro.*

	N	Media	Deviazione std.
CERCA	36619635	7.759	2.600
FORZE	36619635	43.251	1.919
LAUREA	36619635	9.723	2.053
DIPLOMA	36619635	29.680	3.330
TERZIARI	36619635	60.175	7.766

TAB. 1.7b. *Statistiche descrittive delle variabili della tab. 1.6: regioni del Mezzogiorno.*

	N	Media	Deviazione std.
CERCA	20943719	21.776	4.345
FORZE	20943719	35.026	1.547
LAUREA	20943719	9.754	.959
DIPLOMA	20943719	28.876	1.730
TERZIARI	20943719	64.751	3.324

Emerge con chiara evidenza la fortissima differenza esistente tra Nord-Centro e Mezzogiorno per quanto riguarda la disoccupazione:

---

trale (Toscana, Umbria, Marche, Lazio); Italia Meridionale (Abruzzo, Molise, Campania, Basilicata, Calabria); Italia Insulare (Sicilia, Sardegna).

la percentuale di persone in cerca di occupazione nel Sud è quasi tre volte quella del resto d'Italia. Il livello di istruzione, espresso dalla percentuale di laureati e di diplomati, non è invece molto diverso nelle due grandi ripartizioni territoriali.

5. *Le matrici dei dati a tre vie*

Un'importante generalizzazione — non considerata nel Vol. I — è quella delle matrici dei dati a tre «vie» (*three-way data matrices*). In termini generali, esse sono caratterizzate dal fatto che ciascun dato elementare che vi compare, x_{ist}, presenta tre indici, che corrispondono ad una classificazione dello stesso in base a tre criteri (Coppi e Bolasco, 1989; Rizzi (ed.), 1995; Rizzi e Vichi, 1995):

$$\underset{n\times p\times q}{\mathbf{X}} = [x_{ist}] \text{ per } i = 1, ..., n;\ s = 1, ..., p;\ t = 1, ..., q.$$

Esempi tipici sono i seguenti:

i) una matrice dei dati del tipo «unità × variabili × occasioni» (ove il termine «occasioni» può indicare: diversi tempi, differenti situazioni sperimentali, luoghi diversi, etc.); in questo caso la matrice dei dati viene talvolta chiamata «cubica», anche se il termine risulta improprio, poiché si è in presenza d'un parallelepipedo di dati;

ii) una successione di matrici di indici di prossimità (v. cap. IV) del tipo «oggetti × oggetti» rilevate in differenti «occasioni» (ad esempio, la similarità tra n prodotti di marche differenti, valutata dai consumatori in anni successivi, oppure in regioni diverse).

Osservazione I. In base alla definizione generale precedente potrebbe considerarsi come una matrice a tre vie anche una matrice dei dati partizionata, assumendo in tal caso come «occasioni» i gruppi di unità. Infatti, ciascun dato elementare può essere scritto: $x_{is(g)}$, ove g è l'indice del gruppo di appartenenza, con numerosità n_g, e $i = 1, ..., n_g$; $s = 1, ..., p$. Si noti che in questa situazione i gruppi possono avere differente numerosità e questo implica che le matrici a due vie siano di diversa dimensione.

Osservazione II. Nel primo esempio citato i tre criteri in base ai quali si considera il dato sono differenti (unità, variabili e occasioni), mentre nell'esempio *ii*) i primi due criteri sono uguali (oggetti ed oggetti) ed il terzo diverso (occasioni). I criteri di classificazione, ciascuno dei quali genera un insieme di indici, vengono spesso chiamati «modi»,

per cui nell'esempio *i*) la matrice dei dati è a tre vie e tre modi, mentre nel secondo è a tre vie ma a due modi.

Osservazione III. Secondo la definizione generale sopra fornita anche una tabella a tripla entrata potrebbe considerarsi *formalmente* come una matrice a tre vie. In essa però i «dati», n_{ijt}, sono costituiti dalle frequenze (di casella) delle terne di modalità dei tre caratteri esaminati. Riteniamo quindi preferibile mantenere distinto quest'ultimo caso, per la differenza sostanziale esistente tra frequenze e valori.

Le estensioni sopra indicate delle usuali matrici del tipo unità × variabili se da un lato consentono di arricchire l'articolazione (o la profondità) dell'analisi, dall'altro evidenziano la soggettività (e l'ampiezza) della scelta dei criteri in base ai quali si può studiare un determinato insieme di informazioni statistiche. Si tratta, in sostanza, di criteri di classificazione (ad esempio, unità, variabili, occasioni, tempi, modalità d'una generica variabile, etc.) che lo statistico sceglie come dimensioni lungo le quali organizzare la massa delle informazioni a disposizione.

Un'ulteriore generalizzazione è rappresentata dalle matrici dei dati a più di tre vie — in cui ogni dato elementare è indicizzato in base a più di tre criteri — ma questa situazione, oltre che maggiormente complicata, risulta solitamente di minore interesse applicativo.

In questa sede ci limiteremo a considerare un caso particolare, ma estremamente importante, delle matrici a tre vie e precisamente quando esse sono del tipo: unità × variabili × tempi. Nel caso menzionato sulle medesime unità sono rilevate le stesse variabili in più tempi successivi. La matrice di dati a tre vie si presenta allora nella forma seguente:

$$\underset{n\times p\times q}{\mathbf{X}} = \begin{bmatrix} \vdots \\ \cdots\ x_{is1}\ \cdots \\ \vdots \end{bmatrix} \cdots \begin{bmatrix} \vdots \\ \cdots\ x_{ist}\ \cdots \\ \vdots \end{bmatrix} \cdots \begin{bmatrix} \vdots \\ \cdots\ x_{isq}\ \cdots \\ \vdots \end{bmatrix} \tag{1.1}$$

e può essere pensata come una successione temporale di q matrici dei dati $\mathbf{X}_t = [x_{ist}]$, del tipo consueto unità × variabili. Questo corrisponde a sezionare il parallelepipedo dei dati in «fette frontali» (*frontal slices*) (6).

Alcuni esempi sono i seguenti:

i) la rilevazione in n aziende di p variabili desunte dai dati del bilancio in q anni successivi;

ii) la misurazione in n pazienti di p sintomi, esprimibili quantitativamente, in q giorni;

iii) i valori d'un insieme d'indicatori demografici, economici e sociali per ciascuna delle province italiane, rilevati dai Censimenti Istat negli anni 1961, 1971, 1981 e 1991.

I dati del tipo suddetto sono chiamati *longitudinali* e la loro caratteristica essenziale è che le medesime unità statistiche sono «misurate» ripetutamente nel corso del tempo (Diggle *et al.*, 1994). Le analisi longitudinali estendono gli studi basati sull'abituale matrice dei dati unità × variabili (sovente chiamati dati *sezionali* o di tipo *cross-section*), in cui per ogni unità statistica si dispone d'un solo vettore di dati riferito ad un istante (o ad un intervallo temporale) prefissato. L'interesse delle matrici dei dati a tre vie, in cui una dimensione è il tempo, risiede nel fatto che esse permettono di «misurare il cambiamento», con riguardo alle singole unità e ai differenti fenomeni, consentendo un'analisi dinamica (Plewis, 1985; Von Eye, (ed.), 1990).

Un esempio economico molto interessante di matrice a tre vie è fornito dalle azioni quotate in Borsa (le unità statistiche), dalle seguenti variabili: prezzo ufficiale, prezzo di riferimento (cioè calcolato sull'ultimo 10% delle contrattazioni giornaliere), numero di contratti effettuati, controvalore degli scambi, considerando più giorni successivi di Borsa aperta. Questi dati sono riportati giornalmente da «Il Sole - 24 ore» e nelle pagine economiche dei principali quotidiani. Un esempio molto ridotto della suddetta matrice a tre vie è presentato nella tab. 1.8, che considera 10 *blue chips*, due variabili (prezzo ufficiale, in Euro, e controvalore, in migliaia di Euro), in tre giorni consecutivi (29, 30 e 31 marzo 1999).

---

(6) In via alternativa, la matrice a tre indici può essere pensata come:

a) un insieme di p *slices* laterali (matrici di dimensioni $n \times q$, ciascuna delle quali contiene i valori d'una variabile, rilevati per le n unità nei q tempi);

b) un insieme di n *slices* orizzontali (matrici di dimensioni $p \times q$, contenenti i valori delle p variabili nei q tempi per ogni singola unità).

TAB. 1.8. *Prezzi in Euro e controvalore, in migliaia di Euro, di 10 blue chips in 3 giorni consecutivi.*

	TITOLO	PREZZO1	VALORE1	PREZZO2	VALORE2	PREZZO3	VALORE3
1	COMIT	7.588	76522	7.666	86200	7.649	86044
2	ENI	5.713	138004	5.800	197649	5.894	170053
3	FIAT	2.886	26145	2.938	43073	3.029	79270
4	GENERALI	36.541	93513	37.260	113451	37.130	86458
5	INA	2.657	54270	2.708	105850	2.780	139930
6	MEDIASET	8.395	18082	8.540	44043	8.653	24488
7	MONTEDISON	.948	12337	.959	12888	.953	12382
8	OLIVETTI	2.881	22916	2.881	60848	2.916	73644
9	PIRELLI	2.482	14007	2.482	11573	2.572	15855
10	TELECOM	9.744	530002	9.744	652963	9.798	354070

Se le unità statistiche della matrice a tre vie formano un campione estratto da un certo universo, esse costituiscono un *panel.* Le indagini su dati provenienti da un *panel* costituiscono pertanto — secondo la definizione qui accolta — un caso particolare, ma saliente, delle analisi longitudinali (7). Dei tre esempi citati, solo il primo ed il secondo possono ritenersi inerenti a *panels*, purché rispettivamente le n aziende e gli n pazienti costituiscano un campione casuale d'un universo di riferimento.

Da ultimo segnaliamo che «recentemente notevole attenzione è stata rivolta alla possibilità di sfruttare insiemi di informazioni provenienti da una successione di *indagini sezionali ripetute* nel tempo (IRS)» (Rettore e Torelli, 1994). In tal caso le unità statistiche incluse nei campioni in tempi diversi non sono le medesime, per cui l'indagine longitudinale non può essere riferita ai singoli individui (che cambiano nel tempo), bensì a categorie o classi (aggregati di unità elementari), che si conservano immutate per l'intero periodo di osservazione. Ad esempio, si possono classificare le famiglie congiuntamente per numero di componenti ed a seconda della condizione del capofamiglia e per ciascuna di tali n classi si può considerare il vettore delle p voci di spesa in q anni. La matrice risultante è a tre vie del tipo: unità $\times$ variabili $\times$

(7) Le indagini di questo tipo fanno sorgere una serie di problemi — schemi di campionamento, perdita di unità nel corso del tempo, etc. — per i quali rinviamo alla letteratura specifica (Hsiao, 1986; Kasprzyk *et al.*, 1989; il vol. 8, n. 4, 1996 della rivista *Statistica Applicata* ed il vol. 24, n. 2, 1998 della rivista *Survey Methodology*, entrambi dedicati alle indagini longitudinali).

tempi, ma occorre sottolineare che in questo caso le variazioni intertemporali sono attribuibili in parte anche alle fluttuazioni di campionamento, poiché i dati (medi) di ciascuna classe sono desunti da campioni diversi nei vari anni.

Le n unità statistiche considerate in una matrice a tre vie possono anche essere partizionate. Un esempio è costituito dai comuni d'una provincia distinti a seconda della zona altimetrica (pianura, collina e montagna); per ciascuno di essi si possono rilevare le seguenti variabili: numero di nati, di morti, di immigrati e di emigrati in ognuno dei mesi del 1999. L'analisi longitudinale dei dati pone in luce la dinamica delle variazioni naturali e del movimento migratorio per gruppi di comuni appartenenti alla medesima zona altimetrica.

6. *Il trattamento dei dati mancanti*

Nelle analisi statistiche si presenta molto frequentemente la situazione in cui per alcune unità non sono disponibili i valori di una o più variabili. Per un'ampia illustrazione dei problemi suscitati dai dati mancanti (*missing values*) rinviamo al vol. I, pp. 317-327, alla biliografia ivi citata ed inoltre al recente volume di Schafer (1997). In questa sede ci limitiamo a ricordare in breve le scelte più semplici che può effettuare il ricercatore di fronte a tale situazione e che sono abitualmente previste come opzioni nei *packages* statistici.

i) Esclusione *listwise*. Le unità statistiche con anche un solo valore mancante delle p variabili vengono eliminate. Pertanto, le analisi multidimensionali sono sempre riferite al medesimo numero di unità statistiche e precisamente quelle che non hanno alcun dato mancante.

ii) Esclusione *pairwise*. Nelle analisi bidimensionali, per ogni coppia di variabili si eliminano le unità statistiche che hanno un valore mancante per almeno una di tali variabili. Conseguentemente, gli indici statistici bivariati (ad esempio, il coefficiente di correlazione) possono essere calcolati su numeri differenti di unità statistiche per le diverse coppie di variabili.

Torneremo sul problema dei dati mancanti con specifico riferimento a ciascuna delle metodologie multivariate che illustreremo in seguito.

7. *La matrice di covarianza e la matrice di correlazione*

Consideriamo una matrice riferita ad n unità statistiche e p variabili quantitative. Com'è noto (vol. I, pp. 151-161), la relazione lineare tra due generiche variabili quantitative può essere misurata tramite la covarianza o il coefficiente di correlazione.

I valori della covarianza tra ciascuna delle coppie di variabili sono raccolti nella seguente matrice di covarianza, di dimensioni $p \times p$:

$$\mathbf{S} = [\text{cov}(X_s, X_v)] \text{ per } s, v = 1, 2, \ldots, p.$$

Analogamente, i valori dei coefficienti di correlazione sono raccolti nella matrice di correlazione, di dimensioni $p \times p$:

$$\mathbf{R} = [r_{sv}] \text{ per } s, v = 1, 2, \ldots, p.$$

Nel caso generale di ponderazione delle unità statistiche la matrice di correlazione può ottenersi (fra le altre formule) dalla seguente espressione (in funzione della matrice dei dati in termini di scostamenti standardizzati e della matrice diagonale dei pesi relativi):

$$\mathbf{R} = \mathbf{Z}'\mathbf{W}\mathbf{Z} \tag{1.2}$$

Nel caso di unità non ponderate (cioè con pesi relativi tutti uguali a $1/n$) l'espressione precedente si riduce a:

$$\mathbf{R} = \frac{1}{n}\mathbf{Z}'\mathbf{Z}. \tag{1.3}$$

Si tenga presente che sia la matrice di covarianza sia la matrice di correlazione sono semidefinite positive (8).

Esempio. Con riferimento agli indicatori economici delle 10 aziende alimentari, riportati nella tab. 1.4, calcoliamo la matrice di correlazione, dapprima non ponderata e quindi ponderata in base al fatturato. Per il primo caso, scegliendo dal menù la sequenza: *statistica-cor-*

(8) Una matrice simmetrica $\mathbf{A}$ si dice semidefinita positiva (s.d.p.) se e solo se $\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0$ per ogni vettore $\mathbf{x}$ ($\mathbf{x} \neq 0$). Se la disuguaglianza precedente è forte (cioè vale con il segno $>$) la matrice $\mathbf{A}$ si dice definita positiva (d.p.).

relazione-bivariata ed indicando le variabili d'interesse, si ottiene la matrice di correlazione riportata nella tab. 1.9.

TAB. 1.9. *Correlazione non ponderata per le variabili della tab. 1.4.*

	ECON.PRO	CASH.FAT	LAVOR.VA	ROE	INDE.CAP
ECON.PRO					
N					
CASH.FAT	.533				
N	10				
LAVOR.VA	-.266	-.617			
N	10	10			
ROE	.828	.808	-.529		
N	10	10	10		
INDE.CAP	-.688	.007	-.073	-.267	
N	10	10	10	10	

In ogni casella compare, oltre al valore del coefficiente di correlazione lineare, anche il numero, indicato con N, di unità statistiche su cui esso è calcolato. Nel caso di dati mancanti e con l'esclusione *pairwise* tale numero può non essere uguale per tutte le coppie di variabili. Inoltre, scegliendo le opzioni corrispondenti, in ogni casella è possibile ottenere altre informazioni: la significatività del coefficiente di correlazione (che illustreremo nel paragrafo successivo), i prodotti degli scarti e le covarianze (9).

Per il caso ponderato, come si è già detto per il calcolo delle statistiche descrittive, occorre preliminarmente scegliere dal menù l'opzione: *pesa i casi*, indicando come peso il FATTURATO, e quindi ripetere la sequenza precedente. Si ottiene la matrice di correlazione riportata nella tab. 1.10, nella quale N corrisponde alla somma dei pesi.

(9) Lo stesso programma consente anche di calcolare gli indici di cograduazione τ di Kendall e ρ di Spearman (Vol. I, pp. 227-240).

TAB. 1.10. *Matrice di correlazione ponderata per le variabili della tab. 1.4.*

	ECON.PRO	CASH.FAT	LAVOR.VA	ROE	INDE.CAP
ECON.PRO					
N					
CASH.FAT	.544				
N	19203				
LAVOR.VA	-.299	-.623			
N	19203	19203			
ROE	.865	.729	-.504		
N	19203	19203	19203		
INDE.CAP	-.751	-.206	.094	-.453	
N	19203	19203	19203	19203	

Si può osservare che nel caso in esame la matrice di correlazione non ponderata e quella ponderata mostrano valori dei coefficienti di correlazione corrispondenti non molto diversi tra loro. Ad esempio, il legame diretto più stretto si registra in entrambi i casi tra ROE e ECON.PRO (rispettivamente 0.828 e 0.865); similmente, la correlazione inversa più forte è quella tra INDE.CAP e ECON.PRO (–0.688 e –0.751). La concordanza sostanziale tra i risultati è dovuta principalmente al fatto che in questa applicazione i pesi adottati non sono molto diversi fra loro (le aziende considerate sono tutte grandi), e pertanto non vi è una marcata dissomiglianza dalla situazione non ponderata, nella quale si attribuisce implicitamente un peso uguale a $1/n$ a ciascuna unità. Si noti inoltre che per alcune coppie di variabili l'impiego della ponderazione comporta un aumento della correlazione, mentre per altre implica una riduzione.

8. *La significatività della correlazione*

Un problema che il ricercatore si pone di fronte ad una matrice di correlazione è distinguere le relazioni importanti da quelle trascurabili. Da un punto di vista puramente descrittivo, si dice che un coefficiente di correlazione in modulo prossimo ad 1 segnala una relazione lineare quasi perfetta per la corrispondente coppia di variabili, mentre un va-

lore vicino a zero indica l'assenza d'un legame lineare. Meno immediata risulta invece l'interpretazione di valori intermedi, anche se si può affermare che valori più elevati, in modulo, del coefficiente di correlazione segnalano una relazione lineare più stretta (a parità di numero di unità statistiche considerate).

Quando le n unità statistiche rappresentano un campione casuale estratto da un certo universo, è possibile impostare il problema suddetto in termini inferenziali, fornendo una risposta sulla significatività della correlazione, ad un livello prefissato di probabilità (10).

Se le n osservazioni sono determinazioni di due variabili osservate contemporaneamente su un campione di n unità, interessa modellare la corrispondente popolazione in termini d'una variabile aleatoria (v.a.) doppia (X, Y), con funzione di densità $f(x, y)$. In particolare, si vuole applicare una procedura induttiva riguardante il coefficiente di correlazione ρ dell'universo, utilizzando le n coppie (x_i, y_i) di osservazioni campionarie (intese come « realizzazioni » della corrispondente v. a. doppia).

Prima di proseguire occorre specificare la forma di distribuzione ipotizzata per la v.a. doppia (X, Y). Assumiamo che il campione si possa ritenere proveniente da una distribuzione normale bivariata (v. Vol. I, pp. 161-164) con coefficiente di correlazione ignoto ρ (parametro nell'universo). Tale distribuzione ha un ruolo privilegiato nei problemi inferenziali sul coefficiente di correlazione, analogamente alla normale univariata per i problemi sulla media.

Pur con l'assunzione semplificatrice di normalità bivariata, la distribuzione campionaria del coefficiente di correlazione, r, è assai complessa nel caso generale d'un valore qualsiasi del parametro $\rho\,(-1 \leq \rho \leq +1)$. Essa è però più semplice quando $\rho = 0$, risultando solo per tale valore simmetrica rispetto a 0. Infatti, è stato dimostrato che, se $\rho = 0$, la distribuzione campionaria di r è tale che la statistica:

$$t_r = \frac{r}{\sqrt{1 - r^2}} \sqrt{n - 2} \qquad (1.4)$$

ha una distribuzione T di Student con $(n - 2)$ gradi di libertà.

---

(10) In questo paragrafo si presuppone che il lettore conosca gli elementi di base dell'inferenza statistica, al livello di un primo corso universitario di Statistica. Per un'introduzione chiara all'argomento si vedano, ad esempio, Cicchitelli (1992), Piccolo (1998).

Quanto precede consente di sottoporre a verifica l'ipotesi nulla di assenza di correlazione nell'universo:

$$H_0 : \rho = 0,$$

contro l'ipotesi alternativa bilaterale:

$$H_0 : \rho \neq 0.$$

L'ipotesi nulla di assenza di correlazione può essere respinta, al livello di significatività α, se il valore del coefficiente di correlazione calcolato sul campione di n osservazioni, r_c, è tale che:

$$|t_{r_c}| > t_\alpha$$

ove: t_α è il valore in una v.a. T di Student con $(n-2)$ gradi di libertà al quale sono associate le seguenti probabilità: $\Pr(T \leq -t_a) = \Pr(T \geq t_a) = \alpha/2$.

Se l'ipotesi nulla viene respinta, si può affermare che tra le due variabili esiste una correlazione *significativa*, al livello α prefissato (solitamente $\alpha = 0.05$ oppure 0.01). Vogliamo sottolineare che questa conclusione è corretta se sono valide le assunzioni di partenza, cioè quando le n osservazioni costituiscono un campione casuale tratto da un universo con distribuzione normale bivariata.

È possibile calcolare i valori numerici del coefficiente di correlazione, che risultano significativi al livello α, in corrispondenza d'una numerosità campionaria uguale a n. Partendo dalla disuguaglianza:

$$|t_{r_c}| > t_\alpha$$

con alcuni passaggi si ricava che sono significativi i valori:

$$|r_c| > \sqrt{\frac{t_\alpha^2}{(n-2) + t_\alpha^2}} \tag{1.5}$$

Ad esempio, leggendo sulla tavola i valori di t_α, per $\alpha = 0.05$ i coefficienti di correlazione significativi per diverse numerosità campionarie n risultano i seguenti:

$$n = 5;\ |r_c| \geq 0.878$$

$$n = 10;\ |r_c| \geq 0.632$$

$$n = 20;\ |r_c| \geq 0.444$$

$$n = 40;\ |r_c| \geq 0.312$$

$$n = 100;\ |r_c| \geq 0.197.$$

Per $\alpha = 0.01$, si ottiene:

$$n = 5;\ |r_c| \geq 0.958$$

$$n = 10;\ |r_c| \geq 0.765$$

$$n = 20;\ |r_c| \geq 0.563$$

$$n = 40;\ |r_c| \geq 0.402$$

$$n = 100;\ |r_c| \geq 0.256.$$

Osservazione. Il valore in modulo del coefficiente di correlazione campionario che risulta significativo ad un livello α prefissato diminuisce al crescere della numerosità del campione. A parità di n, diminuendo il livello di significatività aumenta la soglia minima del valore, in modulo, che consente di ritenere significativa la correlazione.

Vogliamo porre in evidenza le differenze d'interpretazione del coefficiente di correlazione in ambito descrittivo ed inferenziale. Ad esempio, un valore del coefficiente di correlazione uguale a 0.3 dal punto di vista puramente descrittivo segnala un legame lineare molto scarso. Se però tale coefficiente è calcolato su un campione di 100 osservazioni, la relazione tra le due variabili deve intendersi come *statisticamente significativa*, anche al livello dell'uno %. Si può quindi affermare, con una probabilità di errore minore di 0.01, che tra le due variabili esiste un legame effettivo e non dovuto semplicemente al caso, anche se la relazione è ben lontana da quella perfettamente lineare. Si noti che il medesimo valore numerico $r_c = 0.30$ non risulterebbe significativo neppure al livello del 5% se il campione fosse di 40 unità.

I *packages* statistici principali forniscono, oltre al calcolo del valore numerico del coefficiente di correlazione, anche il cosiddetto *p-value*, definito in generale come la probabilità che la statistica in esame assuma valori in modulo maggiori di quello osservato quando è vera l'i-

potesi nulla. Con riferimento all'ipotesi qui considerata: $H_0 : \rho = 0$, il *p-value* è definito come segue:

$$\Pr\{|t_r| > t_{r_c} \mid H_0 \textit{ vera}\}$$

Pertanto, valori del *p-value* piccoli (minori di 0.05 oppure minori 0.01) portano a rifiutare l'ipotesi nulla. Il *p-value* rappresenta il *livello di significatività osservato* e viene indicato con «Sig.» nell'output di SPSS (cioè *Significantly different from zero*).

Esempio. Nella tab. 1.11 sono riportati i dati ottenuti nel controllo statistico della qualità effettuato da un'azienda alimentare sulle uova utilizzate nella produzione di tagliatelle (fonte dei dati: rilevazione diretta presso l'azienda). Il campione è costituito da 15 prelievi effettuati in maniera casuale sulla materia prima impiegata nel corso della produzione e le variabili rilevate sono le seguenti:

residuo secco, in % (RESIDUO);
colesterolo, in % (COLEST);
lipidi, in % (LIPIDI).

Per accertare se esiste una relazione significativa tra dette variabili, in SPSS si è utilizzata la sequenza: *statistica-correlazione-bivariata*, ottenendo la matrice di correlazione, con indicazione del *p-value*, riportata nella tab. 1.12. Da essa si evince che tra colesterolo e lipidi e tra residuo secco e lipidi la correlazione risulta significativa al livello dell'1% (contro un'alternativa bilaterale, cioè nel test a due code), mentre tra residuo secco e colesterolo la correlazione non è significativa. Infatti, il corrispondente *p-value* è uguale a 0.265, e questo indica che se si rifiutasse l'ipotesi nulla di assenza di correlazione si rischierebbe di commettere un errore di prima specie (rigetto d'una ipotesi nulla vera) con probabilità del 26.5%, molto superiore ai consueti livelli di significatività. La procedura considerata di SPSS fornisce quindi direttamente le risposte cercate. A scopo didattico, mostriamo però interamente lo svolgimento dei calcoli seguendo i passi prima indicati, con riferimento alla relazione tra colesterolo e lipidi. Il coefficiente di correlazione campionario è uguale a 0.747, per cui si ottiene:

$$t_{r_c} = \frac{0.747}{\sqrt{1 - 0.747^2}} \sqrt{15 - 2} = 4.051$$

Sulla tavola della v.a. T di Student per 13 gradi di libertà e per $\alpha = 0.01$ si legge: $t_{0.01} = 3.012$, per cui essendo $t_{r_c} > t_\alpha$, si può rifiutare

l'ipotesi nulla di assenza di correlazione lineare tra le due variabili, accettando quindi l'ipotesi alternativa dell'esistenza d'un legame effettivo tra colesterolo e lipidi.

TAB. 1.11. *Valori del residuo secco, del colesterolo e dei lipidi in un campione di 15 prelievi delle uova impiegate nella produzione di tagliatelle.*

	RESIDUO	COLEST	LIPIDI
1	25.23	.467	10.57
2	25.31	.465	10.52
3	25.38	.467	10.71
4	25.41	.462	10.53
5	25.70	.470	10.83
6	25.57	.455	10.65
7	25.67	.473	11.02
8	25.14	.457	10.51
9	25.22	.456	10.39
10	25.53	.465	10.65
11	25.13	.462	10.52
12	25.34	.473	10.72
13	25.20	.471	10.79
14	25.37	.470	10.68
15	25.26	.462	10.46

TAB. 1.12. *Matrice di correlazione e p-value per i dati della tab. 1.11.*

		RESIDUO	COLEST	LIPIDI
RESIDUO	Correlazione di Pearson			
	Sig. (2-code)			
	N			
COLEST	Correlazione di Pearson	.307		
	Sig. (2-code)	.265		
	N	15		
LIPIDI	Correlazione di Pearson	.676**	.747**	
	Sig. (2-code)	.006	.001	
	N	15	15	

**La correlazione è significativa al livello 0.01 (2-code).

Facciamo notare, infine, che nel caso in esame sono soddisfatte le condizioni per l'impiego della procedura inferenziale sui coefficienti di

correlazione, poiché le osservazioni costituiscono un campione casuale estratto da un universo per il quale è plausibile l'ipotesi di normalità bivariata per le coppie di variabili di volta in volta considerate.

9. *Misure di variabilità multidimensionale*

Le n osservazioni p-dimensionali possono essere pensate (vol. I, p. 141) come punti nello spazio a p dimensioni. Il centroide è il baricentro di tale nuvola di punti e rappresenta la generalizzazione della media nel caso multidimensionale.

È interessante considerare inoltre la dispersione dei punti attorno al centroide, che porta ad introdurre le misure della variabilità multidimensionale (11).

Una prima misura sintetica della variabilità multidimensionale è la somma delle varianze, anche detta *varianza totale*, che corrisponde alla traccia (somma degli elementi sulla diagonale principale) della matrice di covarianza:

$$VAR_T = \sum_{s=1}^{p} \text{var}(X_s)$$

La varianza totale presenta problemi interpretativi se le variabili hanno diversa unità di misura ed inoltre ha il difetto di non tenere conto della correlazione esistente tra le variabili e questo rappresenta indubbiamente un difetto. Infatti, « si può ritenere che la nozione di variabilità corrisponda a quella di *estensione* dell'insieme dei punti raffiguranti i casi osservati e che una misura di un campo minimo entro il quale è contenuto l'insieme o la maggior parte di esso possa rappresentare un indice di variabilità » (Lunetta, 1981, p. 47).

L'estensione effettiva della nuvola di punti nello spazio dipende dalla correlazione tra le variabili. Per comprendere questo concetto, consideriamo il caso più semplice di due sole variabili X e Y. Se esse sono incorrelate, i punti nel diagramma di dispersione tendono a disporsi casualmente nel rettangolo i cui vertici corrispondono alle seguenti 4 coppie di valori:

---

(11) Per un approfondimento di questo tema si vedano: Lunetta (1973), De Carolis (1983) e gli Atti del Convegno SIS, 1981, Pavia - Salice Terme, vol. II, pp. 9-130, dedicati a « Recenti tendenze nello studio della variabilità ».