applica anche a sottoinsiemi non ottenuti con procedura casuale («campioni non probabilistici»);

- in genere si utilizza il simbolo *N* per la numerosità della popolazione ed *n* per la numerosità (o ampiezza) del campione; tuttavia nelle pagine che seguono, poiché non tratteremo l'inferenza campione – popolazione ma ci riferiremo semplicemente al numero totale di casi studiati, utilizzeremo solo il simbolo *N*;
- le caratteristiche delle unità studiate sono dette *proprietà*; ogni proprietà può assumere degli *stati* diversi (per esempio, la proprietà «pratica religiosa» assume gli stati di praticante, saltuario, non praticante);
- la *variabile* è la proprietà operativizzata, cioè rilevata sui casi attraverso una certa procedura detta «definizione operativa»; per esempio possiamo operativizzare la proprietà «pratica religiosa» attraverso la domanda «Nell'ultimo anno lei è andato in chiesa?»;
- le *modalità* sono gli stati della variabile e *valori* i simboli assegnati alle modalità; i valori in genere, anche se non necessariamente, sono numeri. Per esempio, la variabile «pratica religiosa» operativizzata nel modo sopra riportato, ha le seguenti modalità: «mai», «due-tre volte l'anno», «una volta al mese», «due-tre volte al mese», «una o più volte la settimana»; i cui rispettivi valori sono: 1, 2, 3, 4, 5. Se la variabile è nominale gli stati della proprietà vengono anche chiamati «categorie»;
- *variabili dicotomiche* (*dicotomìe*) sono le variabili con due modalità; *variabili politomiche* quelle a più di due modalità;
- l'*analisi monovariata* consiste nell'analizzare le variabili singolarmente prese, cioè ad una ad una senza metterle in relazione fra di loro; l'*analisi bivariata* è lo studio delle relazioni fra due variabili; l'*analisi multivariata* è lo studio delle relazioni intercorrenti fra più di due variabili[3].

2. MATRICE DEI DATI

Supponiamo di aver operativizzato i concetti oggetto dello studio mediante le domande di un questionario. Per esempio, per rilevare un concetto complesso come quello di partecipazione politica ne abbiamo identificati gli indicatori (che, lo ricordiamo, sono ancora dei concetti ma più semplici e

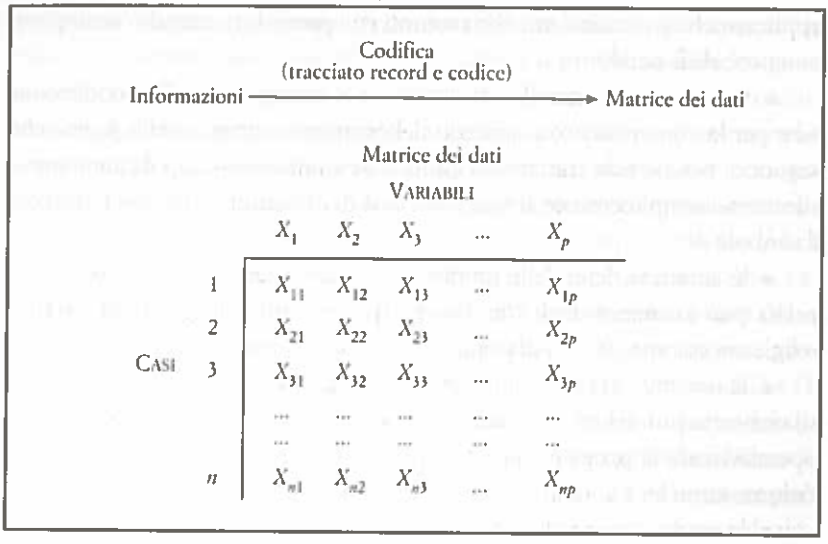

fig. 2.1. *La matrice dei dati (casi × variabili: C × V).*

più facilmente rilevabili empiricamente) che abbiamo operativizzato con delle domande come: «Lei legge il giornale?» oppure «Lei si interessa di politica, molto, abbastanza, poco o per nulla?». Ponendo la domanda e trascrivendo la risposta, trasformiamo il concetto in variabile, cioè operativizziamo il concetto (o, per meglio dire, la proprietà dei soggetti studiati[4]). Abbiamo così ottenuto un pacco di questionari che rappresentano il materiale empirico raccolto.

Si tratta ora di organizzare questo materiale empirico grezzo in una forma tale da poter essere analizzato con gli strumenti dell'analisi statistica. In generale, nella ricerca quantitativa, questo processo di organizzazione del materiale empirico consiste nella sua trasformazione in una matrice di numeri, la *matrice dei dati*, detta anche matrice «casi per variabili» (C × V; cfr. fig. 2.1). La matrice dei dati (d'ora in poi «matrice-dati») consiste in un insieme rettangolare di numeri, dove in riga abbiamo i casi ed in colonna le variabili; in ogni cella derivante dall'incrocio fra una riga e una colonna abbiamo un **dato**, e cioè il valore assunto da una particolare variabile su un particolare caso.

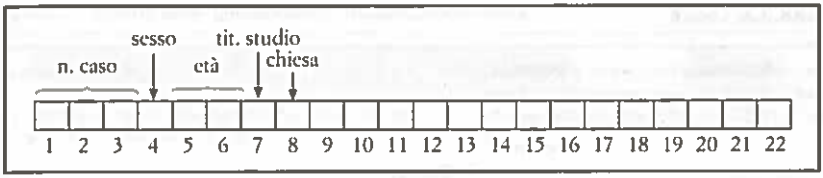

fig. 2.2. *Tracciato record.*

Due sono le condizioni necessarie perché le informazioni afferenti ad un certo insieme di casi possano essere organizzate nella forma di matrice-dati:

- l'unità d'analisi deve essere sempre la stessa: per esempio deve trattarsi di informazioni raccolte tutte su individui, oppure tutte su comuni o su famiglie; non si può fare una matrice-dati dove alcune righe contengono individui, altre comuni, ecc.;
- su tutti i casi studiati devono essere state rilevate le stesse informazioni; nella matrice-dati le righe hanno la stessa lunghezza e contengono le stesse variabili; non è possibile costruire una matrice dati se su un certo numero di casi sono state raccolte certe informazioni e su altri casi ne sono state raccolte delle altre[5].

L'operazione di traduzione del materiale empirico grezzo (il pacco di questionari, la pila di documenti) in matrice-dati viene chiamata **codifica**, ed avviene con l'ausilio di due strumenti, il tracciato-record ed il codice:

- il *tracciato-record* indica la posizione di ogni variabile nella riga della matrice (per esempio dice che la variabile «genere» si trova nella colonna 4 della matrice); il termine «tracciato-record» deriva dalla lingua inglese, dove per «record» si intende la riga della matrice;
- il *codice* (*codebook*) assegna ad ogni modalità della variabile un valore numerico (per esempio dice che nella variabile genere, si assegna il valore 1 a «maschio» e 2 a «femmina»).

In figura 2.2 abbiamo riportato il tracciato-record relativo a quattro variabili derivanti da un ipotetico questionario, e in tabella 2.2 il relativo codice. Nella pratica della ricerca, molto spesso tracciato-record e codice sono incorporati nel questionario stesso. Ne abbiamo riportato un esempio più avanti in tabella 2.3. Come si vede, accanto ad ogni domanda è riportata la

TAB. 2.2. Codice

COLONNA	VARIABILE
1-3	*N. caso*
4	*Genere* 1. Maschio 2. Femmina
5-6	*Età* (riportare l'età in anni)
7	*Titolo di studio* 1. Licenza elementare 2. Licenza media 3. Diploma 4. Laurea 9. Non risponde
8	*Nell'ultimo anno lei è andato in chiesa?* 1. No, mai 2. Due-tre volte l'anno 3. Una volta al mese 4. Due-tre volte al mese 5. Una o più volte la settimana 9. Non risponde

posizione della variabile generata dalla domanda stessa sulla riga (funzione del tracciato-record); ogni alternativa di risposta inoltre è numerata e il numero corrisponde al valore riportato nella matrice (funzione del codice). In figura 2.3 abbiamo riportato la matrice relativa a 10 casi ai quali sia stato somministrato il questionario della tabella precedente.

Ogni riga (record) di questa matrice corrisponde ad un caso (un individuo, un questionario): leggendo una riga sappiamo come quell'individuo ha risposto alle domande (si può dire che ogni riga ci fornisce il *profilo* di un caso). Ogni colonna della matrice corrisponde ad una variabile: leggendo una colonna conosciamo la sequenza di risposte date a quella domanda da tutti gli intervistati. Come si vede dalla figura 2.3, una matrice-dati è un insieme non comprensibile di numeri, che però diventano intelligibili con l'ausilio del tracciato-record e del codice. Prendiamo il numero 4 che compare all'incrocio fra la riga 3 e la colonna 8: si tratta di un dato che appartiene al caso n. 3; il tracciato-record (riportato sul questionario) ci dice che, essendo sulla colonna 8, si riferisce alla variabile «orientamento verso l'inflazione» (domanda 3); il codice (riportato sul questionario) ci dice che il valore 4 corrisponde a «per niente importante». Quindi l'individuo n. 3 alla do-

TAB. 2.3. Estratto da un questionario con codifica incorporata

N. ordine generale |_| |_| |_| |_| 1-4

1. Secondo lei oggi in Italia rispetto a 5 anni fa, si vive meglio, nello stesso modo o peggio?

- Meglio 1 |_| 5
- Stesso modo 2
- Peggio 3
- Non risponde 9

2. Si parla molto di ciò che si dovrebbe fare in Italia nei prossimi 10 o 15 anni. In questo cartellino sono indicati vari obiettivi. (MOSTRARE CARTELLINO 1). Secondo lei quale di questi è il più importante, cioè quale metterebbe al primo posto? E quale al secondo?

	primo	seconda	
• Mantenere l'ordine nel paese	1	1	|_||_|6-7
• Aumentare per i cittadini la possibilità di partecipare al governo	2	2	
• Controllare l'aumento dei prezzi	3	3	
• Garantire la libertà d'opinione	4	4	
• Non risponde	9	9	

3. Vorrei conoscere il suo punto di vista su alcuni problemi sociali. Mi dica per ciascuno di questi se lei lo considera un problema molto, abbastanza, poco o per niente importante?

	abb. mol.			nie. poc.	NR	
• L'inflazione, l'aumento dei prezzi	1	2	3	4	9	|_| 8
• La disoccupazione	1	2	3	4	9	|_| 9
• L'inefficienza dei servizi pubblici	1	2	3	4	9	|_|10
• La criminalità organizzata	1	2	3	4	9	|_|11
• La droga	1	2	3	4	9	|_|12
• L'inquinamento	1	2	3	4	9	|_|13
• L'immigrazione straniera	1	2	3	4	9	|_|14
• La corruzione politica	1	2	3	4	9	|_|15

4. Ora le leggerò un elenco di problemi urbani. Vorrei sapere di quale di questi, qui a ... (NOME DEL COMUNE), lei è maggiormente insoddisfatto (MASSIMO DUE RISPOSTE)

	Ia risp.	IIa risp.	
			|_| |_| 16-17
• Trasporti pubblici	1	1	
• Orari degli uffici pubblici	2	2	
• Traffico	3	3	
• Nettezza urbana	4	4	
• Illuminazione pubblica	5	5	
• Approvvigionamento acqua	6	6	
• Non risponde	9	9	

5. Ora le leggerò alcune azioni che la gente talvolta fa per protestare o per influire sul governo. Lei mi dovrebbe dire se le è mai capitato di compiere qualcuna di queste azioni.

	sì	no-NR	
• Partecipare a scioperi spontanei	1	2	|_|18
• Bloccare il traffico	1	2	|_|19
• Fare l'autoriduzione dell'affitto	1	2	|_|20
• Fare l'autoriduzione delle bollette	1	2	|_|21
• Occupare case sfitte	1	2	|_|22
• Occupare fabbriche	1	2	|_|23
• Scrivere slogan sui muri	1	2	|_|24

6. Nel nostro studio abbiamo incontrato molte persone che, pur andando a votare, dichiarano di non interessarsi affatto di problemi politici. Secondo lei per quali ragioni in genere la gente non si interessa di politica?

|_\|\|_\|25-26

_____

_____

_____

_____

Codifica della domanda 6

00. Non è vero, tutti partecipano
01. Gli italiani sono immaturi
02. Per disinteresse
03. Per particolarismo degli elettori
04. Per particolarismo dei governanti
05. Per disgusto
06. Per paura
07. Perché il sistema crea confusione
08. Perché il sistema è inefficiente
09. Per la disonestà dei governanti
10. Perché la politica è incomprensibile
99. Non risponde

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
0	0	0	1	1	1	3	4	4	2	3	3	1	3	3	6	2	1	1	2	1	2	2	2	0	0
0	0	0	2	3	4	4	3	2	1	2	3	1	3	2	3	3	1	2	1	2	2	2	2	0	9
0	0	0	3	9	9	9	4	4	9	3	3	2	3	2	9	9	2	1	1	1	2	1	1	0	3
0	0	0	4	1	3	3	1	3	4	2	2	1	2	1	2	3	2	1	2	2	1	1	2	0	2
0	0	0	5	1	2	2	1	2	3	2	3	2	4	2	1	2	2	1	1	2	2	2	1	1	0
0	0	0	6	3	2	4	2	3	2	4	2	3	1	3	5	3	2	2	2	1	2	2	1	0	6
0	0	0	7	2	3	1	3	2	2	3	4	2	2	4	2	2	2	2	2	2	2	1	2	0	8
0	0	0	8	9	1	1	3	1	3	9	1	4	9	2	1	6	1	2	2	1	2	2	1	0	0
0	0	0	9	1	4	2	2	2	1	1	1	3	3	3	2	1	1	2	2	2	2	1	1	0	1
0	0	1	0	1	3	3	1	4	2	1	2	4	3	1	3	5	1	2	2	2	1	2	1	0	5

fig. 2.3. *Matrice dei dati (relativa al questionario di tab. 2.3 (10 casi).*

manda se considera l'inflazione un problema importante, ha risposto «per niente importante».

In questo modo abbiamo trasformato il pacco di questionari in una matrice rettangolare di numeri: se i casi sono 1.100 e le variabili 200, si tratterà di una matrice di 1.100 × 200 (1.100 righe per 200 colonne[6]). Tutte le informazioni codificate del questionario si ritrovano nella matrice-dati. La matrice dei dati opera quindi nel senso di una potente organizzazione-concentrazione delle informazioni.

Richiamiamo l'attenzione del lettore su due termini introdotti per la prima volta in questo capitolo – normalmente non tradotti ed utilizzati nella loro formulazione in lingua inglese – che devono diventare familiari all'analista dei dati:

- ogni singola riga della matrice (vale a dire ogni caso trasformato in dati) si chiama *record*;
- la matrice-dati memorizzata su supporto informatico si chiama *file*.

Prima di passare all'analisi dei dati, ci soffermiamo ancora un momento sull'operazione di codifica, cioè di trasformazione delle modalità delle variabili in valori. Analizziamo prima il caso assai comune in cui l'unità d'analisi sia l'individuo ed i dati siano stati raccolti tramite un questionario. Poi vedremo il caso di unità d'analisi differenti dall'individuo.

Scorriamo le domande del questionario in tabella 2.3. La prima domanda («Secondo lei oggi in Italia...») è una domanda standard che non pone problemi e dà luogo ad una variabile. Notiamo solo che fra le risposte codificate, oltre a quelle proposte dalla domanda chiusa, abbiamo aggiunto anche il valore 9 per la modalità di risposta «non risponde». Questa eventualità va sempre prevista in tutte le domande.

La domanda 2 produce due variabili: quella relativa agli obiettivi indicati per primi, e quella relativa agli obiettivi indicati per secondi. La terza domanda è di fatto una *batteria di domande*, nel senso che si tratta di più domande aventi lo stesso formato di risposta (quattro alternative da «molto» a «per niente»). Per comodità grafica esse sono raggruppate apparentemente sotto un'unica domanda, ma di fatto si tratta di 8 domande differenti, che danno luogo a 8 variabili ognuna con la sua sequenza di risposte.

Le due domande 4 e 5 sono *domande a risposta multipla*, e cioè la stessa domanda ammette più risposte. Normalmente la domanda ammette una sola risposta (si faccia attenzione: più alternative di risposta, ma una sola risposta; per esempio nella domanda 1 le alternative erano tre – meglio, stesso modo, peggio – ma la risposta poteva essere una sola). Se la domanda ammette più risposte, la domanda genera più variabili (tante variabili quanto è il numero delle risposte che il soggetto può dare).

La domanda 4 può avere due risposte: un individuo per esempio può rispondere «1» (trasporti pubblici) e «4» (nettezza urbana). Vanno previste quindi due colonne, una per ogni risposta: cioè la domanda genera due variabili (la variabile prima risposta e la variabile seconda risposta). Nella domanda successiva (n. 5) il soggetto può dare tante risposte quante sono le alternative presenti: per esempio uno può teoricamente rispondere che ha partecipato a tutte le forme di protesta elencate. Poiché le alternative proposte sono 7, occorre prevedere 7 colonne, in ciascuna delle quali si registra se il soggetto ha risposto sì o no a quella specifica alternativa. Si noti che anche nel caso precedente si poteva operare in questo modo: si poteva riservare una colonna alla risposta «traffico», registrando sì/no (per esempio 1/2, oppure 1/0) a seconda che il soggetto avesse scelto o meno quel problema; e così via per «trasporti pubblici», «orari degli uffici», ecc.

Infine, la domanda 6 è una *domanda aperta*, cioè non viene offerto all'intervistato un ventaglio di possibili risposte, ma egli risponde come meglio gli

aggrada. In questo caso la codifica viene effettuata a posteriori. In sede di intervista l'intervistatore trascrive sul questionario la risposta liberamente data dall'intervistato; alla fine della rilevazione si leggono tutte le risposte date dagli intervistati (o un consistente campione di esse), si individuano delle categorie in cui classificarle e quindi, sulla base di queste categorie, si codificano le risposte. In tabella 2.3 abbiamo riportato la codifica della domanda 6, per la quale come si vede sono state individuate 10 possibili categorie.

Facciamo ancora due osservazioni. Innanzitutto il numero di cifre delle variabili varia: per esempio la domanda 1 è di una sola cifra (può assumere solo i valori 1, 2, 3, 9), mentre per la domanda 6 si sono riservate nella matrice 2 due colonne, in quanto può assumere valori a due cifre (in questo caso i numeri vanno appoggiati a destra: il valore 1 si codifica 01, il valore 2 con 02, ecc.). Analogamente alla variabile iniziale «numero caso» sono state riservate 4 cifre (sono state intervistate più di 1.000 persone: i casi avranno numero 0001, 0002, ecc.). Inoltre va detto che la matrice può anche contenere dati alfabetici: per esempio avremmo potuto riportare in matrice anche le prime otto lettere del comune di residenza, collocandole (ma è sempre un esempio) nelle colonne 5-12 della matrice, facendo slittare verso destra, a partire dalla colonna 13, il resto del questionario.

Naturalmente l'unità d'analisi può non essere costituita dall'individuo e le informazioni possono essere state raccolte con strumenti diversi dal questionario. L'unità d'analisi può per esempio essere la famiglia, o la scuola, o un'unità produttiva come la fabbrica. Per esempio potremmo avere raccolto una serie di dati sulle scuole medie di un comune, dove le variabili si riferiranno all'unità «istituto scolastico» e potranno essere: quartiere di collocazione della scuola, n. studenti, n. classi, n. insegnanti, n. bocciati, lingue straniere impartite, presenza/assenza di laboratorio linguistico, di palestra, ecc. La struttura della matrice-dati rimane quella vista: sulle righe avremo i casi (cioè gli istituti scolastici) ed in colonna le variabili.

Una situazione di ricerca che si presenta assai di frequente è quella in cui le unità di analisi sono degli **aggregati territoriali** di individui, come i comuni, le province, le regioni. Questo per il semplice motivo che i dati delle fonti statistiche ufficiali si riferiscono a questo tipo di unità. Per far capire al lettore come si configura un *file* di dati aggregati, presentiamo in figura 2.4

COLONNA	VARIABILE
1-2	N. caso
3-6	Nome provincia (prime quattro lettere)
7	Zona geopolitica (da 1 a 4)
8-14	N. abitanti residenti (censimento 1951)
15-21	N. voti validi nel 1953
22-27	N. voti alla Dc nel 1953
28-33	N. voti al Pci nel 1953
34-39	N. analfabeti (cens. '51)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39
0	1	a	l	c	s	1		4	7	7	7	2	2		3	3	1	3	7	8	1	2	0	4	5	2		9	0	6	3	3		1	6	2	1	7
0	2	t	o	r	i	1	1	4	3	3	0	0	1		9	7	4	3	9	5	3	6	1	0	7	2	2	2	9	9	0	5		2	4	0	9	7
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
9	1	s	a	s	s	4		3	4	9	9	5	3		1	8	1	4	8	0		7	7	0	1	7		3	1	3	4	6		5	8	9	9	6

fig. 2.4. *Codice e matrice dei dati nel caso in cui l'unità d'analisi è un aggregato territoriale.*

la parte iniziale di un codice e la relativa matrice-dati di un *file* avente come unità d'analisi la provincia. Come si può vedere non c'è differenza strutturale rispetto ad un *file* di dati individuali. Ci sono delle diversità applicative, in particolare il fatto che le variabili sono per la maggior parte generate da un conteggio (n. abitanti, n. analfabeti, n. ore di sciopero, n. voti al partito A, n. voti validi, ecc.) per cui necessitano di un ampio spazio per la codifica (se la variabile popolazione nella provincia più popolosa, quella di Roma, occupa 7 cifre, a tale variabile dovranno essere riservati 7 campi (caselle) sul codice, e questo per tutte le unità d'analisi, cioè per tutte le province, anche quelle più piccole come Aosta, per la quale la popolazione occupa solo 5 o 6 cifre). Ciò per quanto concerne la configurazione della matrice-dati. Sarà poi nella fase successiva dell'analisi dei dati che queste variabili saranno trasformate (per esempio si dividerà il numero di voti al partito A per il numero di voti validi per avere la percentuale di voti al partito) per rendere possibili i confronti fra le province.

Concludiamo con un cenno sulle procedure di memorizzazione dei dati su supporto informatico o, come talvolta si dice, di «immissione dati» (*data*

entry), che trasferiscono la matrice numerica di figura 2.3 su un supporto di tipo informatico leggibile dal computer. Pur senza soffermarci sull'argomento, ricordiamo che l'operazione può avvenire in maniera assai semplice digitando i valori della matrice sulla tastiera di un computer, creando in questo modo un *file* cosiddetto *Ascii*. Oppure si può utilizzare un foglio elettronico o un *data-base*, soluzione preferibile poiché evita alcuni dei possibili errori di registrazione. Ancora si può ricordare che ci sono procedure automatizzate di immissione dati, come nel caso delle tecniche *Cati* (interviste telefoniche) o *Capi* (interviste faccia-a-faccia), nelle quali il questionario viene letto dall'intervistatore direttamente dal video di un computer e la risposta viene immediatamente digitata su tastiera e memorizzata nella matrice-dati, senza la mediazione di supporti cartacei. Assieme alla matrice, vanno anche fornite al programma di elaborazione (per esempio, il «pacchetto» Spss) le istruzioni di definizione delle variabili (corrispondenti al tracciato-record ed al codice), che permettono al programma stesso di «leggere» la matrice dei dati (si dirà per esempio che nelle colonne 1-3 si trova il numero di identificazione del caso, nella colonna 4 il genere, ecc.).

A questo punto la matrice-dati risulta trasformata nel cosiddetto *system file* (un *file* che incorpora in sé, oltre alla matrice, anche il tracciato-record, il codice, le etichette delle variabili e delle singole modalità di ognuna di esse) ed è pronta per l'analisi statistica.

3. DISTRIBUZIONE DI FREQUENZA

3.1. Distribuzioni assolute e relative

Una volta costruita la matrice-dati, si tratta di analizzarla; analisi che, come abbiamo detto, viene condotta per variabili. Posto di fronte ad una matrice come quella di figura 2.3 (che il lettore immagini come un grande rettangolo che prosegue come minimo per decine di colonne e centinaia di righe) e presa la decisione di cominciare ad analizzare la prima variabile («come si vive oggi in Italia» col. 5, oppure la variabile «titolo di studio»), cosa potrà fare il ricercatore? Come potrà dare una *rappresentazione sintetica* della colonna 5 della matrice? Potrà evidentemente farlo andando a con-