Gestione informatica dei dati

La matrice dei dati

Prof. Roberto Foderà

Cattedra di Gestione informatica dei dati
Dipartimento di Giurisprudenza - Palermo

LUMSA
UNIVERSITÀ

A.A. 2021-2022

La matrice dei dati

I dati possono essere rilevati attraverso un **questionario**, ad esempio chiedendo al cliente i dati rilevanti (nome, cognome, indirizzo, ecc.).
Il questionario è certamente uno strumento efficiente per la raccolta dei dati, scopo per il quale è stato inventato, ma costituisce un supporto inadeguato per la conservazione e, soprattutto, per l'analisi dei dati stessi.
I dati possono essere rilevati anche attraverso **sensori**.

La tecnologia permette di costruire e operare su matrici di dati.

La matrice di dati è una tabella con specifiche relazioni tra righe e colonne.

La matrice dei dati

La matrice di dati, detta anche «**casi per variabili**», consiste in un insieme rettangolare di numeri, dove in riga abbiamo i casi e 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.

$$M_1 = \begin{bmatrix} x_{1,1} & \cdots & x_{1,f} & \cdots & x_{1,p} \\ \vdots & & \vdots & & \vdots \\ x_{i,1} & \cdots & x_{i,f} & \cdots & x_{i,p} \\ \vdots & & \vdots & & \vdots \\ x_{n,1} & \cdots & x_{n,f} & \cdots & x_{n,p} \end{bmatrix}$$

Tratto da Corbetta P., La ricerca sociale, vol. IV, il Mulino, 2003

La matrice dei dati

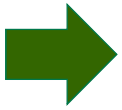
Due sono le condizioni necessarie perché le informazioni afferenti ad un certo insieme di casi possano essere organizzate nella forma di matrice-dati:

• l'unità d'analisi deve essere sempre la stessa: per esempio deve trattarsi di informazioni raccolte tutte su clienti, oppure tutte su ambiti territoriali o, ancora, su prodotti; non si può ottenere una matrice-dati dove alcune righe contengono clienti, altre prodotti ecc.;
nell'esempio l'unità di analisi è la **materia insegnata** in Lumsa.

Insegnamento	Codice Corso	Nome Docente
Gestione inf. dei dati	INFO/01	R. Foderà
Storia economica	SECS-P/12	S. Vinciguerra
Statistica economica	SECS-S/03	G. Notarstefano

La matrice dei dati

• su tutti i casi studiati devono essere state rilevate le stesse informazioni; nella matrice-dati le righe contengono ordinatamente 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.



Ordine delle variabili comune

Insegnamento	Codice Corso	Nome Docente
Gestione inf. dei dati	INFO/01	R. Foderà
Storia economica	SECS-P/12	S. Vinciguerra
Statistica economica	SECS-S/03	G. Notarstefano

Tutti titoli di materie

Tutti codici di materie

Tutti nomi di docenti

La matrice dei dati

Le domande di un questionario possono avere diverse modalità di codifica delle variabili.

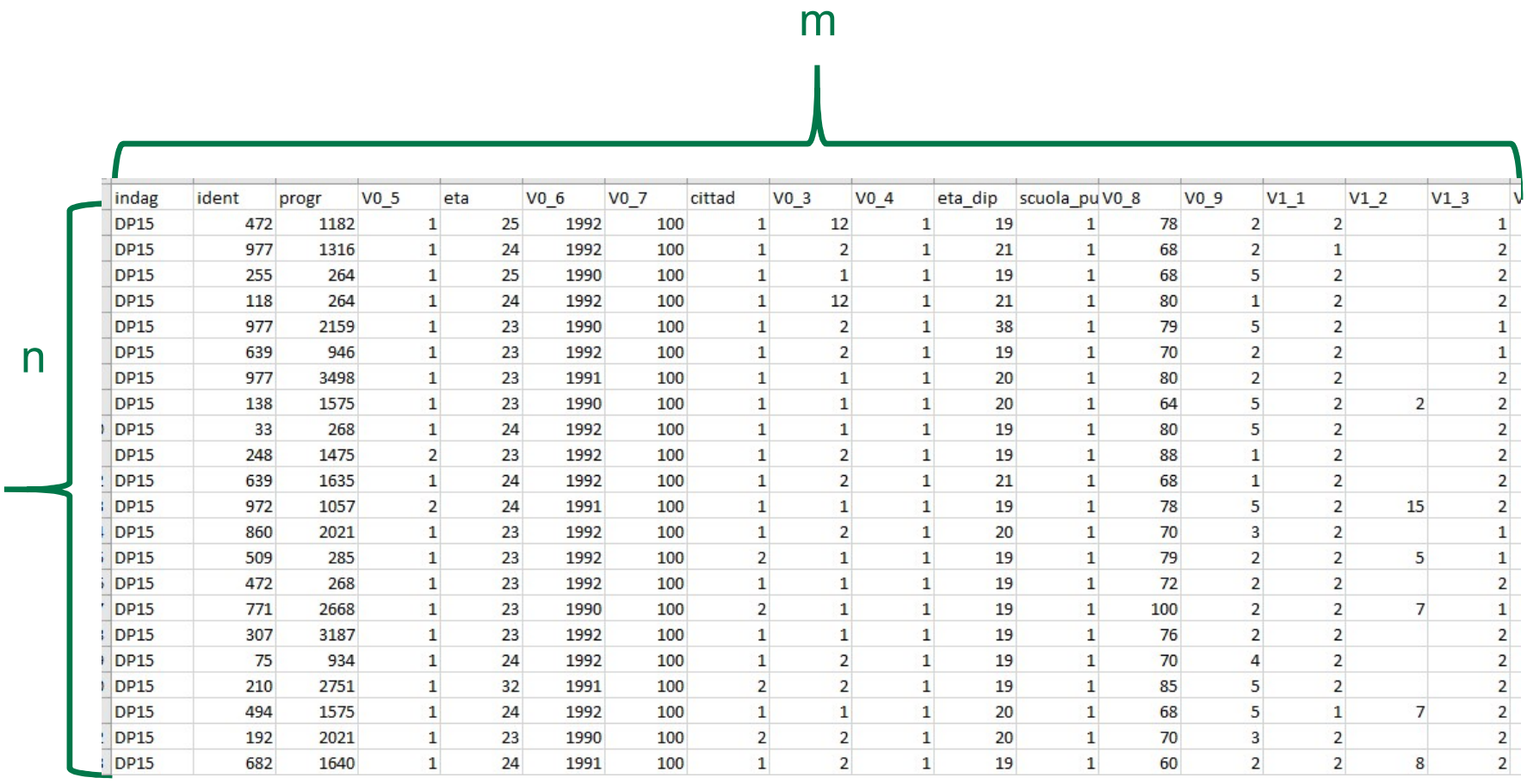
Per tutti

2.3 - Le capita di pensare di avere un carico eccessivo di lavoro domestico?

Leggere le risposte

Sì, spesso 1
Sì, di tanto in tanto 2
Sì, ma raramente 3
No, mai 4

La matrice dei dati



indag	ident	progr	V0_5	eta	V0_6	V0_7	cittad	V0_3	V0_4	eta_dip	scuola_pu	V0_8	V0_9	V1_1	V1_2	V1_3	V
DP15	472	1182	1	25	1992	100	1	12	1	19	1	78	2	2		1	
DP15	977	1316	1	24	1992	100	1	2	1	21	1	68	2	1		2	
DP15	255	264	1	25	1990	100	1	1	1	19	1	68	5	2		2	
DP15	118	264	1	24	1992	100	1	12	1	21	1	80	1	2		2	
DP15	977	2159	1	23	1990	100	1	2	1	38	1	79	5	2		1	
DP15	639	946	1	23	1992	100	1	2	1	19	1	70	2	2		1	
DP15	977	3498	1	23	1991	100	1	1	1	20	1	80	2	2		2	
DP15	138	1575	1	23	1990	100	1	1	1	20	1	64	5	2	2	2	
DP15	33	268	1	24	1992	100	1	1	1	19	1	80	5	2		2	
DP15	248	1475	2	23	1992	100	1	2	1	19	1	88	1	2		2	
DP15	639	1635	1	24	1992	100	1	2	1	21	1	68	1	2		2	
DP15	972	1057	2	24	1991	100	1	1	1	19	1	78	5	2	15	2	
DP15	860	2021	1	23	1992	100	1	2	1	20	1	70	3	2		1	
DP15	509	285	1	23	1992	100	2	1	1	19	1	79	2	2	5	1	
DP15	472	268	1	23	1992	100	1	1	1	19	1	72	2	2		2	
DP15	771	2668	1	23	1990	100	2	1	1	19	1	100	2	2	7	1	
DP15	307	3187	1	23	1992	100	1	1	1	19	1	76	2	2		2	
DP15	75	934	1	24	1992	100	1	2	1	19	1	70	4	2		2	
DP15	210	2751	1	32	1991	100	2	2	1	19	1	85	5	2		2	
DP15	494	1575	1	24	1992	100	1	1	1	20	1	68	5	1	7	2	
DP15	192	2021	1	23	1990	100	2	2	1	20	1	70	3	2		2	
DP15	682	1640	1	24	1991	100	1	2	1	19	1	60	2	2	8	2	

La trasformazione delle risposte dei questionari in un file su un supporto informatico produce una matrice rettangolare n*x*m (n= casi, m=variabili)

La matrice dei dati

L'operazione di traduzione del materiale empirico grezzo (il pacco di questionari, la pila di documenti) in matrice-dati viene chiamata codifica, e 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; 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 o una etichetta.

La matrice dei dati

Esempio di tracciato record

INDAGINE: Iscrizioni e cancellazioni all'anagrafe per trasferimento di residenza Anno 2013					
Tracciato record valido dal 1995					
num. ordine	Posizione	Lunghezza	Acronimovariabile	TipoVariabile	Aggregazione
1	1	4	anno_di_rilevazione	-	
2	5	2	mese_di_rilevazione	Categorica	
3	7	5	filler		
4	12	5	filler		
5	17	1	conteggio	Categorica	
6	18	1	tipo_di_provvedimento	Categorica	
7	19	3	iscrizione_provincia	Categorica	
8	22	3	iscrizione_comune	Categorica	iscrizione_provincia
9	25	3	cancellazione_provincia	Categorica	

La matrice dei dati

Esempio di codebook

ELENCO MODALITA' della variabile: stato_civile (32,0) Anno: 2013

Modalita'	Descrizione	Codice aggregazione	Note
1	celibe/nubile		
2	coniugato/coniugata		
3	vedovo/vedova		
4	divorziato/divorziata		

La matrice dei dati

Le matrici di dati possono essere archiviate e trasmesse secondo diverse tipologie di formati. Quelli maggiormente utilizzati sono:

txt
csv
xsl, xslx
xlm
Html

La matrice dei dati

I sensori e i big dati

La matrice dei dati

Tanti

Non strutturati

Complessi

Non pronti per l’analisi

La matrice dei dati

La costruzione dei dati

Un dato statistico deve essere descritto da alcuni elementi chiave, resi evidenti in ogni corso di statistica, e codificati nel Generic Statistical Information Model (GSIM)

- P Popolazione/unità di riferimento del dato
- V Variabile statistica
- f Operatore statistico

La matrice dei dati

Le variabili statistiche

Preso un collettivo di interesse (popolazione) si studiano uno o più fenomeni sulle singole unità che compongono la popolazione:

1) Questi fenomeni non si manifestano in modo identico su tutte le unità della popolazione e quindi prendono il nome di variabile;

2) Si è soliti distinguere le variabili rispetto al tipo di valore che si associa sulle unità. In generale si distinguono le variabili **qualitative** e categoriali dalle variabili **quantitative** o numeriche;

3) Una variabile **qualitativa** può essere «statisticata» attraverso l'operazione di conteggio.

4) Una variabile **quantitativa** permette di «subire» operazioni matematiche più complesse