L'analisi esplorativa dei dati

- *Introduzione*
- *Osservazioni e dati*

Introduzione

La fase esplorativa d'una ricerca è quella nella quale si fa una prima raccolta di dati empirici, necessari a cercare nella realtà una possibile soluzione alla domanda che ci si è posta.

Esempio 1: Se si studiano gli effetti delle piogge acide sulle foreste, si cercheranno informazioni sull'acidità delle piogge, sulla composizione delle foreste, il loro clima, il tipo d'ambiente umano circostante, ecc.

Esempio 2: Se si vuol sapere la relazione fra pazienti, sintomi d'una malattia e farmaci, si raccoglieranno informazioni relative questi aspetti del problema.

Obiettivi d'un'analisi esplorativa dei dati sono la ricerca di possibili *fattori* che influenzano il fenomeno e una possibile *classificazione* delle osservazioni in gruppi omogenei.

Esempio 1: Nel caso delle piogge acide, l'acidità della pioggia è un fattore, che causa la malattia delle piante. Anche il tipo d'insediamenti umani circostanti è un fattore, perché potrebbe provocare acidità della pioggia. Diverse classi di foreste servono a decidere su quali intervenire.

Esempio 2: Lo stato di salute generale d'un paziente può esser un fattore che favorisce una malattia. Diverse condizioni possono favorire un'evoluzione diversa della malattia e suggerire trattamenti diversi.

Osservazioni e dati

Quando si rilevano elementi relativi ad un aspetto del fenomeno che si studia, si dice che si fanno delle *osservazioni.* Gli elementi rilevati sono degli *attributi* delle osservazioni, che si chiamano *caratteri.*

Esempio 1: un osservazione può esser un rilievo delle piante della foresta, i caratteri sono allora le specie presenti, o la loro abbondanza, l'altitudine, il tipo di suolo, la quantità di pioggia in un mese, ecc.
Esempio 2: Ogni paziente osservato in una data precisa è un'osservazione; i caratteri sono i sintomi osservati, alcune misure (peso, età, lo stato fisico, la febbre, le medicine assunte), ecc.

Sinonimi: osservazioni, individui, unità statistiche;
caratteri, variabili, indicatori statistici.

La *modalità* secondo la quale un dato carattere si presenta in un'osservazione prende il nome di *dato*.

La *tavola di dati* è l'insieme dei caratteri osservati nel corso d'una sperimentazione. Normalmente si fa in modo che in tutte le osservazioni si rilevino gli stessi caratteri, secondo gli stessi criteri.

La tavola s'organizza in modo che ad ogni osservazione corrisponda una riga della tavola e ad ogni carattere corrisponda una colonna.

	Car 1	*Car 2*	*Car 3*	.....	.....	*Car p*
Oss 1	x_{11}	x_{12}	x_{13}		x_{1j}	x_{1p}
Oss 2	x_{21}	x_{22}	x_{23}		x_{2j}	x_{2p}
.....						
Oss i	x_{i1}	x_{i2}	x_{i3}		x_{ij}	x_{ip}
.....						
Oss n	x_{n1}	x_{n2}	x_{n3}		x_{nj}	x_{np}

x_{ij} è la modalità assunta dal carattere *j* nell'osservazione *i*.

I nomi *Car 1*, *Car 2*, *Car 3*, ..., *Car p*, *Oss 1*, *Oss 2*, ..., *Oss n* sono le *etichette*, identificative di caratteri ed unità.

Esempio 1: la tavola dei dati delle foreste

	Altezza	*Orientamento*	*Pendenza*	*Quercus pub.*	....
ril 1	1315	*SW*	12%	4	....
ril 2	915	*N*	0%	3	....
ril 3	1225	*NE*	5%	2	....
.....					

Esempio 2: la tavola dei dati dei pazienti

	Età	*Febbre*	*Sangue*	*Mal di testa*	*Aspirina*	*Colesterolo*	
Rossi	43	38.2	0	si	1	112	...
Letti	35	36.8	A	no	2	-----	...
Magri	83	39.2	B	si	0	155	...
.....							

Dati dall'Annuario Statistico Italiano, Istat, 1978

Regione	*Popolaz.*	*Camere*	*Primar.*	*Second.*	*Terziario*	*Reddito*	*Matrim.*	*Nascite*
Piemonte e Valle d'Aosta	4465.500	6081.300	212	945	723	1721	24.620	50.101
Lombardia	8496.683	9861.440	164	1897	1461	1277	50.166	105.810
Trentino Alto Adige	847.226	1080.889	52	100	167	987	5.332	10.936
Veneto	4135.960	5258.836	206	703	697	1053	27.897	52.820
Friuli Venezia Giulia	1245.143	1722.637	39	178	245	1134	6.813	1.227
Liguria	1868.065	3034.254	52	211	394	1292	9.535	15.943
Emilia Romagna	3852.833	5107.119	278	642	740	1245	21.139	38.352
Toscana	3502.541	4850.555	137	589	636	1126	20.184	36.149
Umbria	773.195	960.612	45	117	141	966	4.951	9.246
Marche	1350.974	1805.806	118	244	230	989	8.553	16.826
Lazio	4764.149	5440.567	172	423	1033	1109	29.862	65.626
Abruzzo	1120.770	1473.219	100	120	198	876	7.877	15.034
Molise	299.775	397.862	51	26	41	706	2.036	3.875
Campania	4984.677	4507.638	394	456	794	816	39.921	96.683
Puglia	3498.932	3381.271	416	315	523	893	27.163	68.134
Basilicata	560.057	559.093	83	55	71	754	4.022	8.812
Calabria	1861.537	1870.928	179	148	256	710	13.550	32.409
Sicilia	4575.421	4802.599	364	370	661	842	31.908	77.726
Sardegna	1441.284	1637.154	87	127	236	954	11.374	25.799

Si riconoscono almeno i seguenti tipi di dati:

- *Dicotomici*: tipo presenza / assenza, di essi si può solo constatare se in un'osservazione si manifestano o no.
- *Qualitativi*: i caratteri presentano modalità differenti, senza alcuna relazione fra di esse.
- *In scala*: le modalità sono dotate d'un ordine totale.
- *Quantitativi discreti*: le modalità sono dei numeri interi, in quantità limitata.
- *Frequenze*: la modalità corrisponde al numero d'elementi che sono rilevati.
- *Quantitativi continui*: vere e proprie *misure*.

A volte esistono *dati mancanti*, se non è stato possibile rilevarli.

Il problema dei dati mancanti non va trascurato. Esistono diverse ragioni per avere dati mancanti e vanno tenute in conto.

Per esempio, è diverso che il dato non sia possibile rilevarlo, oppure che sia possibile ma non sia stato rilevato, oppure che in quella particolare situazione il dato non abbia senso.

Ancora, nei sondaggi, esistono persone che non rispondono mai e che costituiscono un insieme ignoto, sul quale si possono fare solo congetture.