L'ANALISI MONOVARIATA

Tipi di variabili e analisi statistica

- Sono le caratteristiche logico-matematiche delle variabili che definiscono le procedure da seguire nella fase di analisi dei dati.

TAB. 2.1. Tipi di variabili

TIPO DI VARIABILE	OPERAZIONI FRA I VALORI	MISURE DI TENDENZA CENTRALE	MISURE DI DISPERSIONE
Nominale	=	Moda	Indice di omogeneità
Ordinale	><	Mediana	Differenza interquartile
Cardinale	+−×:	Media	Deviazione standard

Nota: Le caratteristiche riportate sono cumulative (per esempio, per le variabili ordinali sono possibili le operazioni di = ≠ > <; per le variabili cardinali si possono calcolare moda, mediana e media; ecc.).

- l'analisi **monovariata** consiste nell'analizzare le variabili singolarmente prese, cioè a una a 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.

Matrice dei dati

- Processo di organizzazione del materiale empirico che consiste nella sua **trasformazione in una matrice di numeri, la matrice dei dati**, detta anche matrice «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**

- L'operazione di traduzione del materiale empirico grezzo (il pacco di questionari, la pila di documenti) in matrice-dati viene chiamata **codifica**

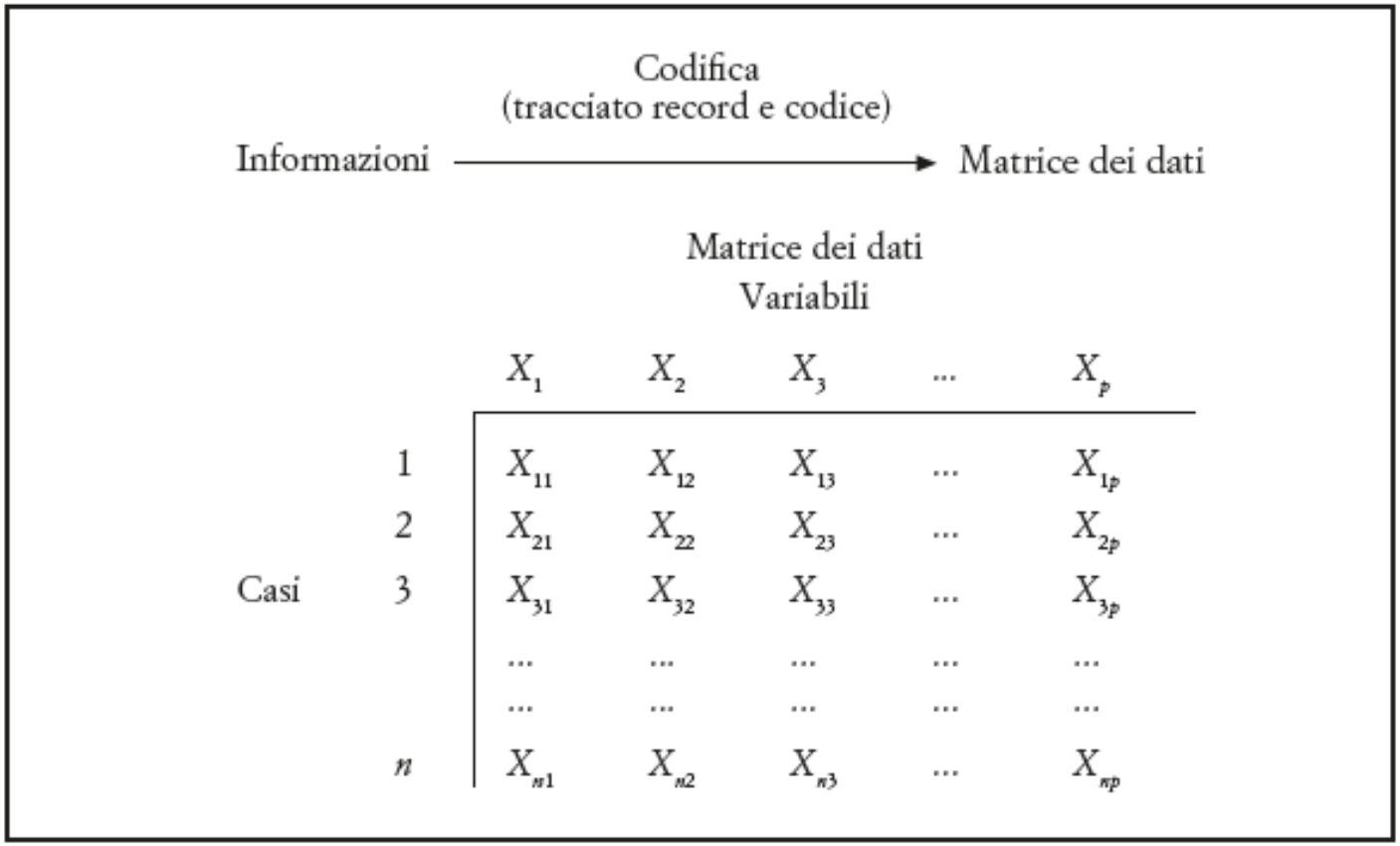


fig. 2.1. *La matrice dei dati (casi × variabili: C × V).*

Distribuzione di frequenza

- **Distribuzioni assolute e relative**
- La distribuzione di frequenza di una variabile è una rappresentazione nella quale a ogni valore della variabile viene associata la frequenza con la quale esso si presenta nei dati analizzati
- Può essere data in forma tabellare, grafica oppure algebrica

- **Distribuzione di frequenza in forma tabellare**

TAB. 2.4. Distribuzione di frequenza della variabile «titolo di studio»

	Frequenze assolute[a]	Frequenze relative		Frequenze cumulate
		Proporzioni	Percentuali	
Senza titolo	30	0,025	2,5	2,5
Licenza elementare	509	0,424	42,4	44,7
Licenza media	342	0,285	28,5	73,4
Diploma	264	0,220	22,0	95,4
Laurea	55	0,046	4,6	100,0
Totale	1.200	1	100,0	

[a] Dette anche valori assoluti (v.a.).

- *Frequenze assolute*: cioè il numero di casi per ciascuna categoria della variabile.
- *Frequenze relative - proporzioni*: posto uguale a 1 il totale dei casi del campione, per ogni categoria viene riportata la proporzione dei casi che appartengono a quella categoria (rispetto al totale pari a 1)
- *Frequenze relative - percentuali*: posto uguale a 100 il totale dei casi del campione, per ogni categoria viene riportata la percentuale dei casi che appartengono a quella categoria (rispetto al totale pari a 100)
- *Frequenze cumulate percentuali*: per ogni categoria viene riportata la percentuale di casi che appartiene a quella categoria e a quelle di grado inferiore. La frequenza dell'ultima categoria e sempre 100% (questo tipo di frequenze può essere usato soltanto con variabili ordinali e cardinali, in quanto richiede che le modalità siano ordinabili).

TAB. 2.5. Distribuzioni di frequenza assolute e relative della variabile «Partito votato alle elezioni per la Camera del 1996, parte proporzionale» in Lombardia e in Emilia-Romagna

	Valori assoluti (in migliaia)		Valori percentuali	
	Lombardia	Emilia-R.	Lombardia	Emilia-R.
Forza Italia	1.510	451	23,6	15,1
Alleanza nazionale	575	344	9,0	11,5
Ccd-Cdu	298	144	4,6	4,8
Lega Nord	1.636	216	25,5	7,2
Pds	965	1.065	15,1	35,7
Lista Dini	267	116	4,2	3,9
Ppi	398	238	6,2	8,0
Verdi	152	75	2,4	2,5
Rifondazione comunista	437	249	6,8	8,3
Altri	168	90	2,6	3,0
Totale	6.406	2.988	100,0	100,0

- E le variabili cardinali?
- Le variabili cardinali come l'età mal si prestano ad essere rappresentate in tabella a causa dell'elevato numero di categorie. Per cui nella distribuzione di frequenza si raggruppano i dati in categorie

TAB. 2.6. Distribuzione di frequenza di una variabile cardinale (distribuzione per età degli operai di uno stabilimento): valori singoli e raggruppati in classi

Età	v.a.	Classi d'età	v.a.
15	1		
16	2		
17	3	15-20	32
18	7		
19	7		
20	12		
21	10		
22	12		
23	12	21-25	72
24	17		
25	21		
26	...	26-30	96
...		31-35	112
...		36-40	130
		41-45	138
		46-50	159
		51-55	142
		56-60	107
		61-65	83

- **E le domande a risposta multipla?**

TAB. 2.7. Presentazione in tabella di distribuzioni di frequenza di domande a risposta multipla

a) Due possibilità di risposta

Domanda: «Di quale dei seguenti problemi lei è maggiormente insoddisfatto?».

	Ia RISP.	IIa RISP.	Ia + IIa RISP.	% SUI RISPONDENTI[a]
Trasporti pubblici	213	12	225	16,1
Orari uffici	322	105	427	30,6
Traffico	557	235	792	56,8
Nettezza urbana	143	43	186	13,3
Illum. pubblica	84	25	109	7,8
Approvv. acqua	75	43	118	8,5
Totale	1.394	463		***
Non risposte	106			

b) Alternativa «Sì/No» a ogni problema

Domanda: «Le è mai capitato di compiere qualcuna di queste azioni?»

	% SÌ
Scioperi spontanei	8,9
Blocco traffico	7,7
Autorid. affitto	6,8
Autorid. bollette	7,1
Occupare case	8,9
Occupare fabbriche	15,3
Slogan sui muri	33,8

[a] Persone che hanno indicato il corrispondente problema su 100 che hanno risposto; la somma delle percentuali supera il valore di 100 in quanto si poteva dare più di una risposta.

Pulizia» dei dati e preparazione del file di lavoro

- Prima di iniziare qualsiasi elaborazione:
  - Controlli di plausibilità
  - Controlli di congruenza
  - Ponderazione
  - Valori mancanti («missing values»)

- Valori mancanti («missing values»)

TAB. 2.9. Distribuzione di frequenza della variabile «pratica religiosa» (Domanda «Nell'ultimo anno lei è andato in chiesa? (se sì) Ogni quanto?»)

a) Tabella di lavoro			*b) Tabella di presentazione dati*		*c) Tabella di presentazione dati*	
		V.A.		%		%
No, mai	1	132	No, mai	8,8	No, mai	9,4
2-3 volte l'anno	2	416	2-3 volte l'anno	27,9	2-3 volte l'anno	29,5
1 volta al mese	3	167	1 volta al mese	11,2	1 volta al mese	11,8
2-3 volte al mese	4	233	2-3 volte al mese	15,6	2-3 volte al mese	16,5
1 volta la settimana	5	415	1 volta la settimana	27,8	1 volta la settimana	29,5
Più volte la settimana	6	35	Più volte la settimana	2,3	Più volte la settimana	2,5
Altra religione	7	11	Altra religione	0,7	Altra religione	0,8
	8	5	Non risponde	5,7		
Non risponde	9	86				
Totale		1.500	Totale	100,0	Totale	100,0
			(N)	(1.495)	(N)	(1.409)
			Valori mancanti: 5		Valori mancanti: 91	

Analisi monovariata

- Misure di tendenza centrale
  - Variabili nominali: **la moda**
  - Variabili ordinali: **la mediana**
  - Variabili cardinali: **la media aritmetica**

- Misure di variabilità
  - Variabili nominali: **indici di omogeneità/eterogeneità**
  - Variabili ordinali: **la differenza interquartile**
  - Variabili cardinali**: deviazione standard e varianza**

Deviazione standard e varianza

- Scostamento semplice medio
- Deviazione standard (scarto quadratico medio)
- Varianza
- Coefficiente di variazione

Varianza

- Il quadrato della deviazione standard è la varianza della distribuzione:

- Varianza $S^2 = \frac{\Sigma(X - \bar{X})^2}{N}$

- La varianza è una misura di grandissima importanza nella statistica. Tutta l'analisi dei dati ruota attorno al concetto di «**varianza spiegata**»: data la variazione di una variabile fra i casi (per esempio, l'orientamento politico che varia fra le persone, il tasso di suicidio che varia fra le nazioni, ecc.) l'analista si chiede con quali altre variazioni di variabili tale variazione è associata.
- Spesso, «spiegare la varianza» della variabile significa anche risalire (sia pure in maniera corroborativa e non dimostrativa) al **meccanismo di causa-effetto** che ha prodotto la sua variazione.
- La varianza quindi, proprio in quanto esprime la variabilità di una variabile, **costituisce l'oggetto primario di tutta l'analisi dei dati**.

TAB. 2.16. Misure di variabilità per variabili cardinali

a) Prima distribuzione

X_i	$(X_i - \overline{X})$	$(X_i - \overline{X})^2$
18	–3	9
20	–1	1
20	–1	1
20	–1	1
21	0	0
23	2	4
25	4	16
$\overline{X} = 21$	$\Sigma(X_i - \overline{X}) = 0$	$\Sigma(X_i - \overline{X})^2 = 32$
	$\Sigma\lvert X_i - \overline{X}\rvert = 12$	

b) Seconda distribuzione

X_i	$(X_i - \overline{X})$	$(X_i - \overline{X})^2$
3	–18	324
6	–15	225
9	–12	144
16	–5	25
20	–1	1
30	9	81
63	42	1.764
$\overline{X} = 21$	$\Sigma(X_i - \overline{X}) = 0$	$\Sigma(X_i - \overline{X})^2 = 2.564$
	$\Sigma\lvert X_i - \overline{X}\rvert = 102$	

Campo di variazione 25 – 18 = 7

Scostamento semplice medio $ssm = \frac{12}{7} = 1{,}7$

Deviazione standard $S = \sqrt{\frac{32}{7}} = \sqrt{4{,}57} = 2{,}1$

Varianza $S^2 = 4{,}57$

Campo di variazione = 63 – 3 = 60

Scostamento semplice medio $ssm = \frac{102}{7} = 14{,}6$

Deviazione standard $S = \sqrt{\frac{2.564}{7}} = \sqrt{366} = 19{,}1$

Varianza $S^2 = 366$

La concentrazione:

- Quando la variabile è cardinale e consiste in quantità possedute dalle unità di analisi, allora si può calcolare la concentrazione di questa variabile nelle unità studiate
- *Rapporto di concentrazione di Gini*
- *Curva di Lorenz*

Rappresentazioni grafiche della distribuzione di frequenza

- Rappresentazioni grafiche di distribuzioni di frequenza di variabili nominali
  - Diagrammi a barre
  - Diagrammi di composizione

Diagrammi a barre

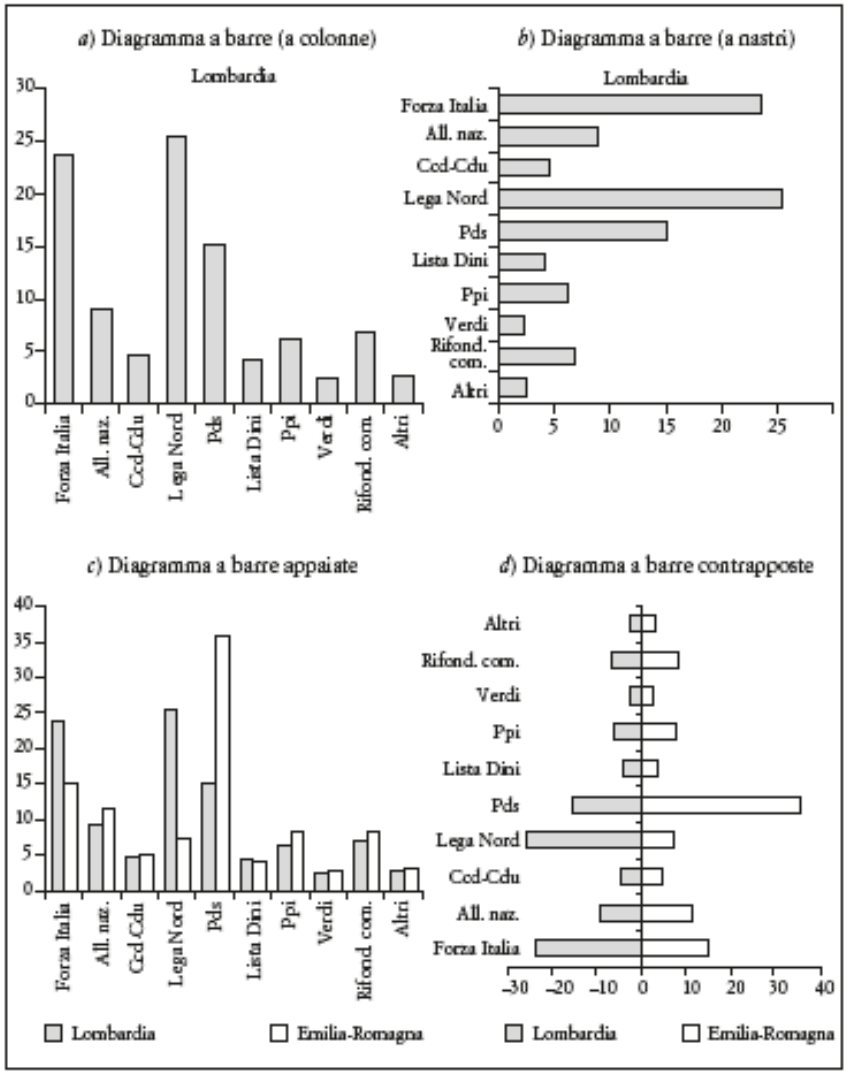


fig. 2.8. *Diagrammi a barre (voto per la Camera nel 1996; sui dati di tab. 2.5).*

Diagrammi di composizione

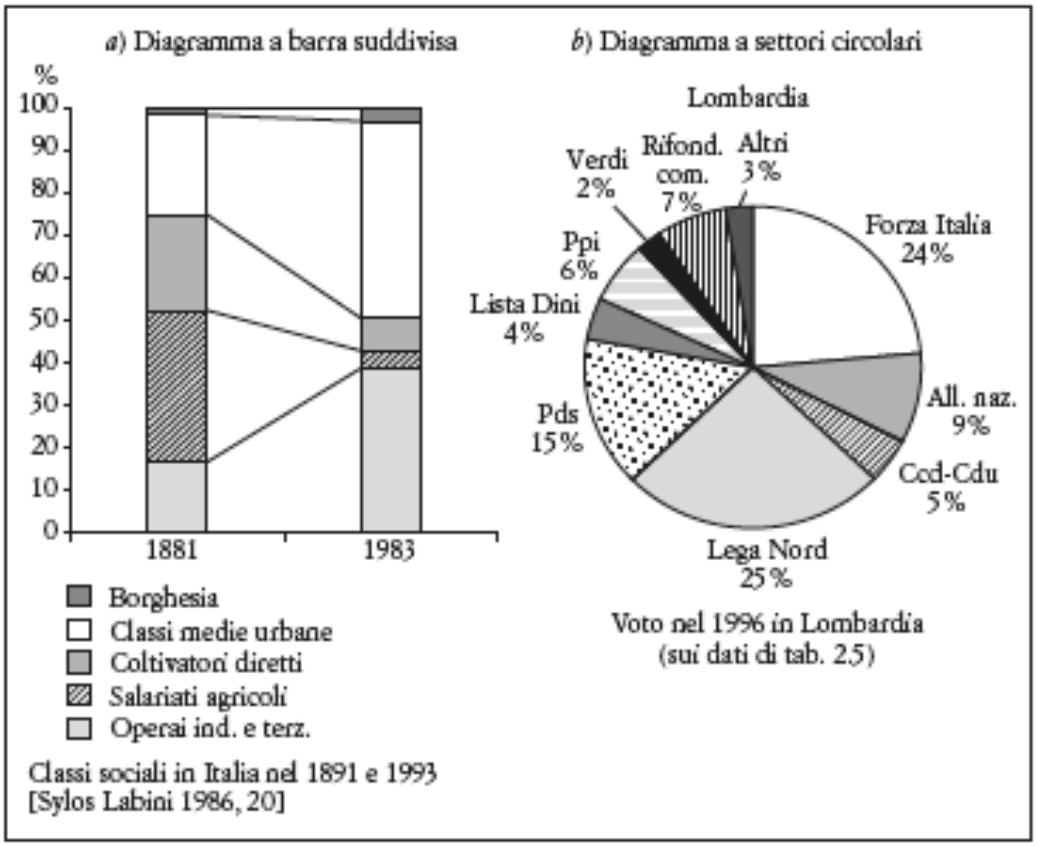
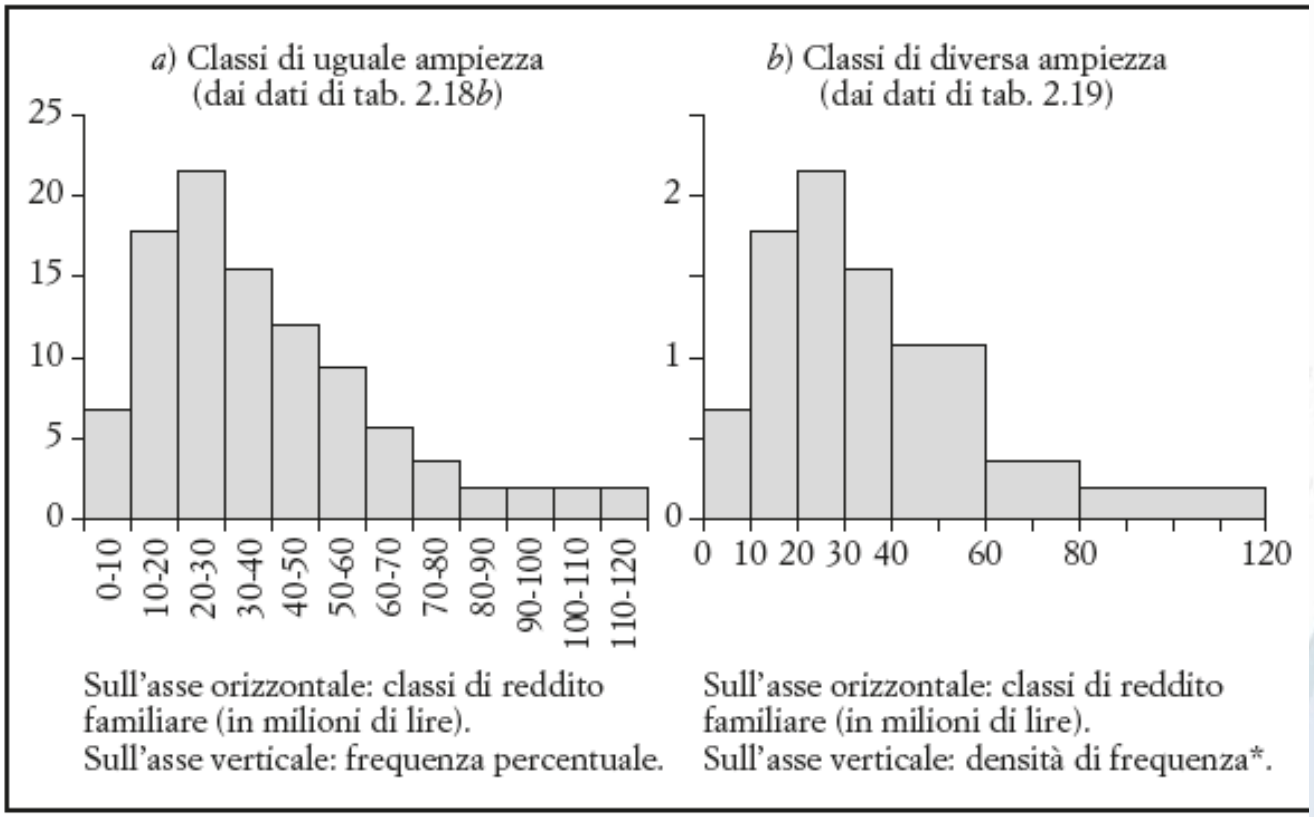


fig. 2.9. *Diagrammi di composizione.*

- Rappresentazioni grafiche di distribuzioni di frequenza di variabili cardinali
  - Istogramma
  - Poligono di frequenza

Istogramma (distribuzione dei redditi delle famiglie italiane nel 1993: reddito annuo in milioni di lire).



Sull'asse orizzontale: classi di reddito familiare (in milioni di lire).
Sull'asse verticale: frequenza percentuale.

Sull'asse orizzontale: classi di reddito familiare (in milioni di lire).
Sull'asse verticale: densità di frequenza*.

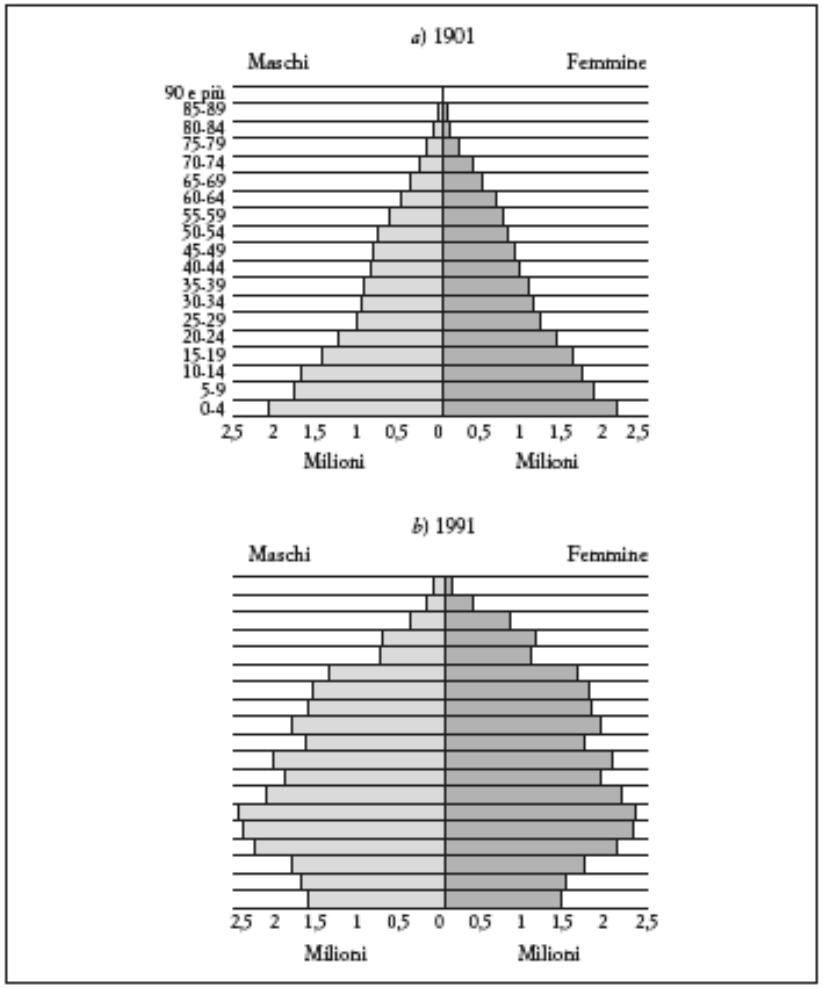


fig. 2.11. *Piramidi delle età della popolazione italiana.*

Fonte: Golini [1994, 156].

Indici di distanza e di dissimilarità

- Indici di distanza fra casi
- Indici di dissimilarità fra distribuzioni

TAB. 2.20. Indici di distanza e di dissimilarità

a) Distanza fra due casi

Matrice-dati

	1	2	3	4	5	6	7
i	3	5	8	2	7	4	5
j	4	7	7	3	6	4	3

$$D_{ij} = \sqrt{(3-4)^2+(5-7)^2+(8-7)^2+(2-3)^2+(7-6)^2+(4-4)^2+(5-3)^2} = 3,4$$

b) Distanza fra un caso e la media di tutti i casi

	Partito A	Partito B	Partito C	Partito D	Totale
Regione *i*	23,5	4,2	38,7	33,6	100
Media nazionale	20,7	6,3	39,1	36,9	100

$$D_i = \sqrt{(23,5-20,7)^2+(4,2-6,3)^2+(38,7-39,1)^2+(33,6-36,9)^2} = 4,82$$

c) Dissimilarità fra distribuzioni di frequenza

	1951	1961
Laurea	1,7	2,4
Diploma	3,7	5,5
Licenza media	6,3	10,8
Licenza elementare	64,5	67,1
Senza titolo	23,8	14,2
Totale	100,0	100,0

$$D = \sqrt{(1,7-2,4)^2+(3,7-5,5)^2+\ldots+(23,8-14,2)^2} = 11,09$$

d) Indice di cambiamento elettorale fra due elezioni

	Ia elezione	IIa elezione
Partito A	25,3	28,2
Partito B	4,2	3,2
Partito C	33,8	34,7
Partito D	36,7	33,9
Totale	100,0	100,0

$$C = \frac{|25,3-28,2|+|4,2-3,2|+|33,8-34,7|+|36,7-33,9|}{2} = 3,8$$

Classificazioni, tipologie e tassonomie

- Nell'analisi dei dati intendiamo per classificazione quel processo secondo il quale i casi studiati vengono raggruppati in sottoinsiemi («classi») sulla base della loro similarità
- I casi possono essere classificati sulla base della loro similarità/dissimilarità su una o più variabili
- Classificazione unidimensionale: aggregazione delle modalità in classi
- Classificazione multidimensionale: tipologie e tassonomie

Trasformazioni delle variabili

- Nel corso dell'analisi statistica le variabili possono essere trasformate per facilitare l'analisi stessa.
- La **standardizzazione** consiste nell'attribuzione alla variabile di nuovi valori, di modo che la media sia pari a 0 e la deviazione standard pari ad 1
- La **normalizzazione** ha lo stesso obiettivo di rendere confrontabili punteggi derivati da variabili misurate con scale diverse, riconducendoli tutti alla scala 0-1
- Gli **indici** sono variabili funzioni di altre variabili, aventi il fine di sintetizzare in un'unica nuova variabile (l'indice) le informazioni contenute in più variabili

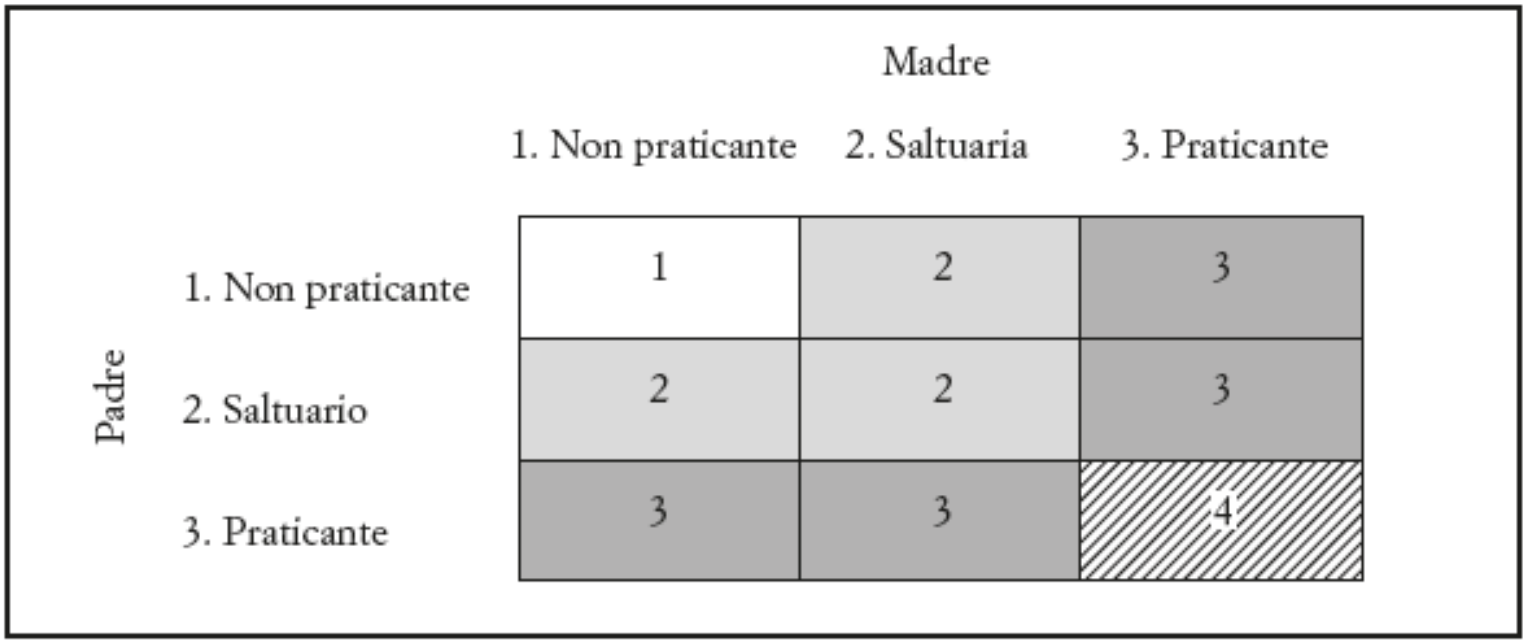


fig. 2.20. *Costruzione dell'indice di «religiosità familiare» sulla base della pratica religiosa del padre e della madre.*

Legenda:
1: assenza di elementi religiosi;
2: debole presenza di elementi religiosi;
3: presenza significativa di elementi religiosi, ma in contraddizione fra i genitori;
4: forte religiosità familiare.

Dati individuali e dati aggregati

- Dati individuali: unità d'analisi è l'individuo
- Dati aggregati: l'unità d'analisi è un aggregato di individui (es. comune, regione, nazione...)
- Problema dei dati aggregati: diversa dimensione dell'aggregato, per cui i dati vanno relativizzati alla dimensione dell'aggregato

- **Rapporti statistici**
- **Rapporti di derivazione**. Si tratta di rapporti fra la misura di un fenomeno e quella di un altro che può essere considerato un suo presupposto necessario.
  - laureati / iscritti all'università;
  - pensioni / popolazione;
  - operai cassaintegrati / totale operai;
  - suicidi / popolazione;
  - reati / popolazione

- **Rapporti medi**. Sono diffusissimi e si hanno tutte le volte che il fenomeno posto al numeratore si può associare mediamente a ogni unità posta al denominatore.

  – rendimento medio per ettaro = tonnellate di grano prodotto / ettari coltivati;

  – densità della popolazione = n. abitanti / superficie del territorio (interpretabile

  come numero medio di abitanti per kmq);

  – indice di affollamento = n. componenti la famiglia / n. stanze dell'abitazione

  (interpretabile come n. medio di persone per stanza);

Serie temporali e territoriali

- **Serie temporale** (o serie storica): sequenza dei valori assunti da una variabile nello stesso aggregato territoriale in tempi diversi
- **Serie territoriale:** sequenza dei valori assunti da una variabile nello stesso momento in aggregati territoriali diversi

TAB. 2.23. Andamento dei divorzi in Italia dal 2000 al 2012 (valori assoluti)

Anno	Divorzi	Incremento percentuale	N. indice (2000 = 100)	N. indice a base mobile
2000	37.573			
2001	40.051	6,6	107	107
2002	41.835	4,5	111	104
2003	43.856	4,8	117	105
2004	45.097	2,8	120	103
2005	47.036	4,3	125	104
2006	49.534	5,3	132	105
2007	50.669	2,3	135	102
2008	54.351	7,3	145	107
2009	54.456	0,2	145	100
2010	54.160	–0,5	144	99
2011	53.806	–0,7	143	99
2012	51.319	–4,6	137	95
Media	47.980			
Deviazione standard	5.850			

Fonte: Istat.

TAB. 2.24. Divorzi per 100 matrimoni per regione, nel 2007

Regione	Divorzi per 100 matrimoni	N. indice (Italia = 100)
Valle D'Aosta	37,6	186
Piemonte	32,3	160
Lombardia	28,5	141
Liguria	33,1	164
Trentino-Alto Adige	26,9	133
Veneto	22,6	112
Friuli Venezia Giulia	30,2	150
Emilia-Romagna	29,8	147
Toscana	26,1	129
Umbria	15,3	76
Marche	20,4	101
Lazio	22,0	109
Abruzzo	17,7	88
Molise	13,7	68
Campania	8,5	42
Puglia	9,6	47
Basilicata	6,9	34
Calabria	9,0	44
Sicilia	13,2	65
Sardegna	15,8	78
Italia	20,2	100

Fonte: Istat.

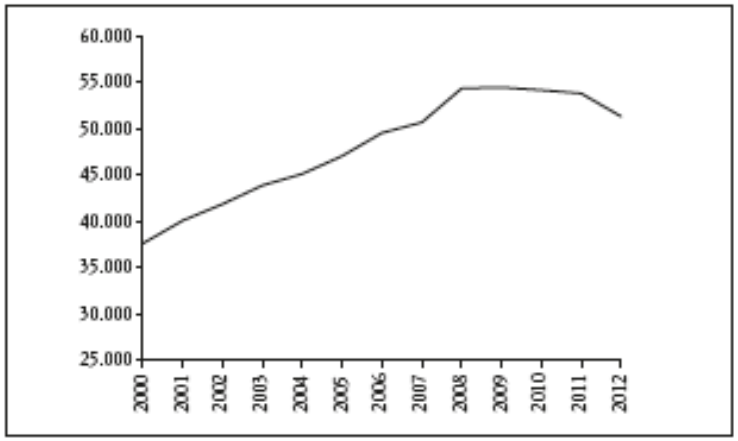


fig. 2.21. *Serie storica: divorzi in Italia dal 2000 al 2012.*

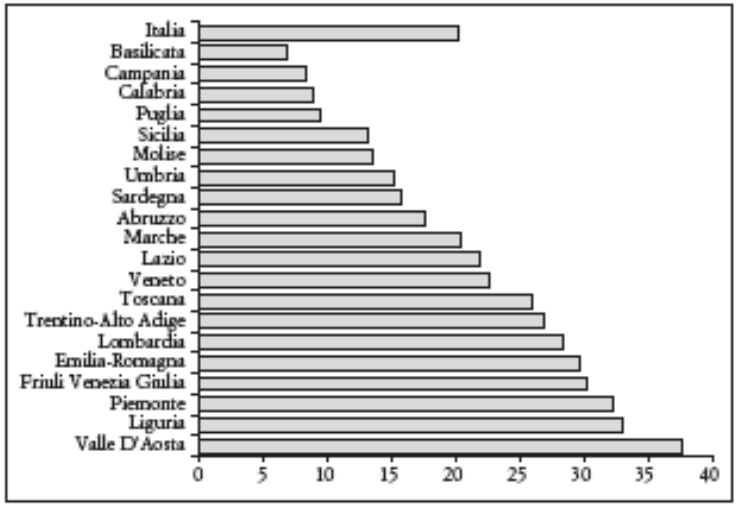


fig. 2.22. *Serie territoriale: divorzi per 100 matrimoni in Italia nel 2007 per regione.*

Lo studio della variazione

- **Differenza assoluta e differenza relativa**
- Per esempio, la variazione del numero di divorzi dai 37.573 del 2000 ai 40.051 del 2001può essere così espressa:

Variazione assoluta: $b - a = 40.051 - 37.573 = +2478$

Variazione relativa: $\frac{b-a}{a} \cdot 100 = \frac{2.478}{37.573} \cdot 100 = 6,6$

- Si dirà dunque che nel periodo considerato i divorzi in Italia sono cresciuti del 6,6%.

- **Numeri indice**. Il numero indice, nell'esempio precedente, permette di rispondere alla seguente domanda: «Se ponessimo uguale a 100 i divorzi nel 2000, a quanto essi ammonterebbero nel 2001?». Si tratta di fare una proporzione

$$37.573 : 100 = 40.051 : X$$

$$X = \frac{40.051}{37.573} \cdot 100 = 106,6$$

- Quindi il numero di divorzi fra il 2000 e il 2001 è passato da 100 a 106,6