L'ANALISI BIVARIATA

Relazioni fra variabili

- Studio delle relazioni fra variabili. Cosa significa dire che c'è una relazione fra due o più variabili?
- Significa dire che c'è una «variazione concomitante» fra i loro valori, una covariazione: per esempio, al variare del titolo di studio varia il reddito

1. Si tratta di relazioni statistiche, cioè relazioni di tipo probabilistico

2. La statistica ci può dire solo che esiste una relazione fra due variabili→Sarà compito e responsabilità del ricercatore di conferire a tale relazione il significato di nesso causale e di attribuire a essa una direzione

TAB. 3.1. Le tecniche di analisi bivariata

		VARIABILE INDIPENDENTE	
		Nominale	*Cardinale*
VARIABILE DIPENDENTE	*Nominale*	Tavole di contingenza	
	Cardinale	Analisi della varianza	Regressione e correlazione

- **Le tavole di contingenza**
  - Percentuali di riga
  - Percentuali di colonna
  - Percentuali sul totale

TAB. 3.2. Pratica religiosa per età

	18-34	35-54	Oltre 54	Totale
a) Tabella dei valori assoluti (frequenze) di cella				
Praticanti	223	313	182	718
Saltuari	266	317	88	671
Non praticanti	425	504	168	1.097
Totale	914	1.134	438	2.486
b) Tabella delle percentuali di riga				
Praticanti	31,1	43,6	25,3	100,0
Saltuari	39,6	47,2	13,1	100,0
Non praticanti	38,7	45,9	15,3	100,0
c) Tabella delle percentuali di colonna				
Praticanti	24,4	27,6	41,6	
Saltuari	29,1	28,0	20,1	
Non praticanti	46,5	44,4	38,4	
Totale	100,0	100,0	100,0	
d) Tabella delle percentuali sul totale				
Praticanti	9,0	12,6	7,3	28,9
Saltuari	10,7	12,8	3,5	27,0
Non praticanti	17,1	20,3	6,8	44,1
Totale	36,8	45,6	17,6	100,0

Fonte: Itanes 1996.

- Si sceglie la percentuale di colonna quando si vuole analizzare l'influenza che la variabile posta in colonna ha sulla variabile posta in riga
- Si sceglie la percentuale di riga quando si vuole analizzare l'influenza che la variabile posta in riga ha sulla variabile posta in colonna
- Si definisce qual è la variabile indipendente e si percentualizza all'interno delle sue modalità

TAB. 3.4. Relazione fra pratica religiosa e comportamento di voto nel 2013

a) Voto a seconda della pratica religiosa (come votano le persone classificate secondo la loro pratica)

	PRATICANTE	SALTUARIO	NON PRATICANTE
Centro-sinistra	26,4	23,6	33,4
Centro-destra	31,7	34,8	26,2
Movimento 5 stelle	19,8	25,9	27,8
Centro	17,3	10,6	7,7
Altri	4,8	5,1	4,9
Totale	100,0	100,0	100,0
(N)	(208)	(216)	(507)

b) Pratica religiosa a seconda del voto (come sono religiosamente connotati gli elettori dei vari partiti)

	PRATICANTE	SALTUARIO	NON PRATICANTE	TOTALE	(N)
Centro-sinistra	20,0	18,5	61,5	100	(275)
Centro-destra	24,1	27,4	48,5	100	(274)
Movimento 5 stelle	17,2	23,5	59,3	100	(238)
Centro	36,7	23,5	39,8	100	(98)
Altri	21,7	23,9	53,4	100	(46)

Fonte: Itanes.

- Presentazione delle tavole
  - Parsimoniosità
  - Totali
  - Basi delle percentuali
  - Cifre decimali, decimale zero, arrotondamenti, quadratura
  - Intestazione

- Interpretazione delle tavole
  - Selezione delle modalità significative della variabile dipendente
  - Errori comuni nell'interpretazione della tabella
  - Aggregazione delle modalità della variabile dipendente
  - Indice di differenza percentuale
  - Forma della relazione

- **Tavole di mobilità sociale**

TAB. 3.14. Tavole di mobilità sociale: classe sociale delle persone occupate in Italia nel 1985 secondo la classe sociale del padre

Classe attuale	Classe d'origine						
	Borghesia	Classe media impiegatizia	Piccola borghesia urbana	Piccola borghesia agricola	Classe operaia urbana	Classe operaia agricola	Totale
a) Valori assoluti							
Borghesia	47	43	35	16	38	4	183
Classe media impiegatizia	54	134	168	83	245	12	697
Piccola borghesia urbana	10	38	221	83	160	39	552
Piccola borghesia agricola	4	0	6	115	0	6	130
Classe operaia urbana	13	38	145	214	500	105	1.014
Classe operaia agricola	0	0	6	10	0	39	55
Totale	128	252	581	521	944	205	2.631
b) Percentuali per colonna							
Borghesia	37	17	6	3	4	2	
Classe media impiegatizia	42	53	29	16	26	6	
Piccola borghesia urbana	8	15	38	16	17	19	
Piccola borghesia agricola	3	0	1	22	0	3	
Classe operaia urbana	10	15	25	41	53	51	
Classe operaia agricola	0	0	1	2	0	19	
Totale	100	100	100	100	100	100	
c) Percentuali per riga							
Borghesia	26	23	19	9	21	2	100
Classe media impiegatizia	8	19	24	12	35	2	100
Piccola borghesia urbana	2	7	40	15	29	7	100
Piccola borghesia agricola	3	0	4	88	0	5	100
Classe operaia urbana	1	4	14	21	49	10	100
Classe operaia agricola	0	0	11	19	0	71	100
d) Percentuali sul totale							
Borghesia	2	2	1	1	1	0	
Classe media impiegatizia	2	5	6	3	9	0	
Piccola borghesia urbana	0	1	8	3	6	1	
Piccola borghesia agricola	0	0	0	4	0	0	
Classe operaia urbana	0	1	6	8	19	4	
Classe operaia agricola	0	0	0	0	0	1	
Totale							100

Immobili (diagonale) = 39
Mobili ascendenti (triangolo superiore) = 37
Mobili discendenti (triangolo inferiore) = 18

Fonte: Cobalti e Schizzerotto [1994].

- **Rappresentazioni grafiche della relazione fra due variabili nominali**
- Si utilizzano gli strumenti già visti per le distribuzioni di frequenza, e cioè sostanzialmente i diagrammi a barre oppure quelli a linee spezzate che congiungono i punti di interesse.

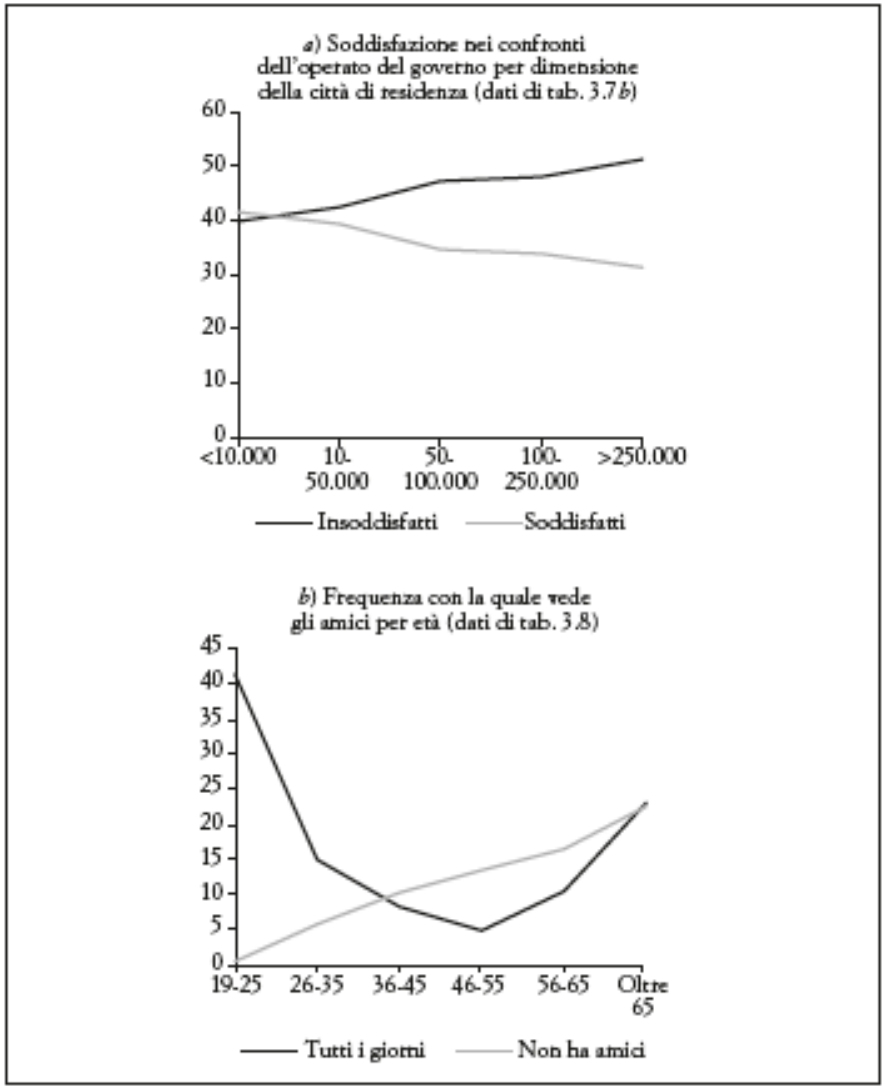

fig. 3.2. *Rappresentazioni grafiche di tavole di contingenza: spezzata.*

- **Significatività della relazione fra due variabili nominali: il test statistico del chi-quadrato (χ^2)**
- Test statistico di verifica delle ipotesi applicato al caso della relazione fra due variabili: formulare l'ipotesi nulla H0 secondo la quale nella popolazione non esiste relazione fra le due variabili e dimostrare, dati alla mano, che essa è falsa: cioè che questa ipotesi non è compatibile (= è assai improbabile) con i dati di cui disponiamo.
- Se l'ipotesi nulla H0 di assenza di relazione viene respinta, automaticamente resta accettata la sua alternativa, l'ipotesi di ricerca H1 che sostiene l'esistenza della relazione.

TAB. 3.15. Frequenze osservate e frequenze attese sotto l'ipotesi nulla H_0 di indipendenza

	18-34		35-54		Oltre 54		Totale	
	v.a.	%	v.a.	%	v.a.	%	v.a.	%
a) Frequenze osservate								
Praticanti	223	24,4	313	27,6	182	41,6	718	28,9
Saltuari	166	29,1	317	28,0	88	20,1	671	27,0
Non praticanti	425	46,5	504	44,4	168	38,4	1.097	44,1
Totale	914	100,0	1.134	100,0	438	100,0	2.486	100,0
b) Frequenze attese sotto l'ipotesi di indipendenza								
Praticanti	264,0	28,9	327,5	28,9	126,5	28,9	718	28,9
Saltuari	246,7	27,0	306,1	27,0	118,2	27,0	671	27,0
Non praticanti	403,3	44,1	500,4	44,1	193,3	44,1	1.097	44,1
Totale	914	100,0	1.134	100,0	438	100,0	2.486	100,0

Calcolo della frequenza attesa per la cella (1,1): $f_e = 914 \cdot 718/2.486 = 264$

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e} = \frac{(223-264,0)^2}{264,0} + \frac{(313-327,5)^2}{327,5} + \frac{(182-126,5)^2}{126,5} + \ldots + \frac{(168-193,3)^2}{193,3} = 45,47$$

$p < 0,001$ (χ^2 significativo al livello dello 0,001);
$\Phi = 0,14$;
$V = 0,10$.

Fonte: Itanes.

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$$

- Per convenzione, noi respingiamo l'ipotesi nulla di indipendenza se $p \leq 0,05$, cioè se il valore del chi-quadrato è così grande da avere solo il 5% (o meno) di probabilità di essere dovuto al caso (cioè a errori casuali pur derivando da una popolazione dove c'è effettiva indipendenza)
- Tavola di distribuzione del χ 2→In questa tavola abbiamo tante righe, cioè distribuzioni del χ 2, quanti sono i gradi di libertà della tabella.
- I gradi di libertà della tabella si determinano nel seguente modo:

gradi di libertà gl = (n. righe – 1) (n. colonne – 1)

Misure della forza della relazione fra variabili nominali e ordinali

- Misure di associazione fra variabili nominali
  - Misure di associazione basate sul chi-quadrato
  - Misure di associazione basate sulla riduzione proporzionale dell'errore
- Misure di cograduazione fra variabili ordinali

- **Rapporti di probabilità (odds)**: rapporto fra la frequenza di una categoria e la frequenza della categoria alternativa (nel caso di variabili dicotomiche)
- lo indichiamo con la lettera greca omega (ω).
- è anche definibile come il rapporto fra la probabilità che un individuo, estratto a caso dall'universo, appartenga a una categoria della variabile considerata e la probabilità che non vi appartenga (da cui il suo nome italiano di «rapporto di probabilità»)

Rapporto di probabilità (*odds*):

$$\omega = \frac{f_1}{f_2} = \frac{p_i}{1 - p_i}$$

TAB. 3.20. Proporzioni e rapporti di probabilità (relazione fra istruzione e atteggiamento verso la pena di morte)

		Istruzione		Totale
		Inferiore	*Superiore*	
Atteggiamento	*Favorevoli*	*a* 1.027	*b* 161	1.188
	Contrari	*c* 397	*d* 207	604
	Totale	1.424	368	1.792

Proporzione: $p = \frac{1.188}{1.792} = 0,663$ Rapporto di probabilità: $\omega = \frac{1.188}{604} = 1,97$

Proporzioni condizionate: $p_1 = \frac{1.027}{1.424} = 0,721$ $p_2 = \frac{161}{368} = 0,438$

Rapporti di probabilità condizionati: $\omega_1 = \frac{1.027}{397} = 2,59$ $\omega_2 = \frac{161}{207} = 0,77$

Rapporto di associazione (*odds ratio*): $\frac{\omega_1}{\omega_2} = \frac{1.027 \cdot 207}{397 \cdot 161} = 3,3$

Fonte: Elaborazione su dati tratti da Corbetta e Parisi [1983].

- **Analisi della varianza**
- (detta anche Anova) serve per studiare la relazione fra una variabile nominale e una cardinale
- Anche in questo caso si può stabilire la significatività della relazione (col rapporto F) e misurarne la forza (con l'eta-quadrato)

- **Teorema fondamentale della varianza**

$$\sum_i \sum_j (Y_{ij} - \overline{Y}_{..})^2 = \sum_i \sum_j (Y_{ij} - \overline{Y}_{.j})^2 + \sum_i \sum_j (\overline{Y}_{.j} - \overline{Y}_{..})^2$$

Somma totale dei quadrati	=	Somma interna dei quadrati (devianza non spiegata)	+	Somma esterna dei quadrati (devianza spiegata)

- Devianza (o somma dei quadrati, SQ)
- Abbiamo così scomposto la devianza della variabile cardinale dipendente in due componenti:
  a) la somma dei quadrati degli scarti dei singoli valori dalla rispettiva media di gruppo; essa viene chiamata somma interna dei quadrati («interna» in quanto è interna al gruppo);
  b) la somma dei quadrati degli scarti delle medie di gruppo dalla media generale, che viene chiamata somma esterna dei quadrati.

- La prima somma è una misura della variabilità del fenomeno entro i gruppi
- La seconda una misura della variabilità del fenomeno studiato fra i gruppi.
- La somma interna dei quadrati viene anche chiamata devianza non spiegata
- La somma esterna viene chiamata devianza spiegata. Spiegata da che cosa?
- Spiegata dalla variabile nominale: è quella parte di variabilità della variabile dipendente che è attribuibile alla variabile indipendente.

$$SQ_{totale} = \underbrace{SQ_{interna}}_{=0 \text{ in caso di relazione perfetta}} + \underbrace{SQ_{esterna}}_{=0 \text{ in caso di assenza di relazione}}$$

Significatività della relazione

- Sottoporre a verifica l'ipotesi nulla secondo la quale le medie di gruppoY provengono tutte da una stessa popolazione e quindi i dati nella popolazione (ipotetica o effettiva) dalla quale derivano sono uguali fra loro

Dividendo la devianza per i gradi di libertà si ottiene la stima della varianza della popolazione dalla quale derivano i dati del campione studiato. I gradi di libertà sono i seguenti:

$$\underset{\text{gradi di libertà totali}}{N-1} = \underset{\text{gradi di libertà interni}}{(N-k)} + \underset{\text{gradi di libertà esterni}}{(k-1)}$$

Le stime della varianza (dette anche «quadrati medi», *mean squares*) sono pertanto:

$$\text{Stima entro i gruppi o stima interna} = \frac{\sum_i \sum_j (Y_{ij} - \overline{Y}_{.j})^2}{N-k}$$

→Se l'ipotesi nulla è vera le due stime sono uguali; se l'ipotesi nulla è falsa la seconda stima è maggiore della prima

TAB. 3.25. Tabella riassuntiva dell'analisi della varianza

	SQ: somma dei quadrati	*GL*: gradi di libertà	Stima della varianza (quadrati medi)	*F*
Totale	2.901,84	$N-1=27$		
Esterna (fra i gruppi, spiegata)	1.979,41	$k-1=3$	659,80	
Interna (entro i gruppi, non spiegata)	922,93	$N-k=24$	38,46	
				17,16

Calcoli:
Stime della varianza: esterna = 1.979,41/3 = 659,80
interna = 922,23/24 = 38,46
Rapporto *F*: = 659,80/38,46 = 17,16

$$\text{Stima fra i gruppi o stima esterna} = \frac{\sum_i \sum_j (\overline{Y}_{.j} - \overline{Y}_{..})^2}{k-1}$$

- Forza della relazione
- eta-quadrato o η^2: rapporto fra la somma dei quadrati esterna (spiegata) e la somma dei quadrati totale (devianza totale):

$$\eta^2 = \frac{SQ_{esterna}}{SQ_{totale}} = \frac{SQ_{spiegata}}{SQ_{totale}}$$

Regressione e correlazione

- Relazione fra due variabili cardinali
- Diagramma di dispersione

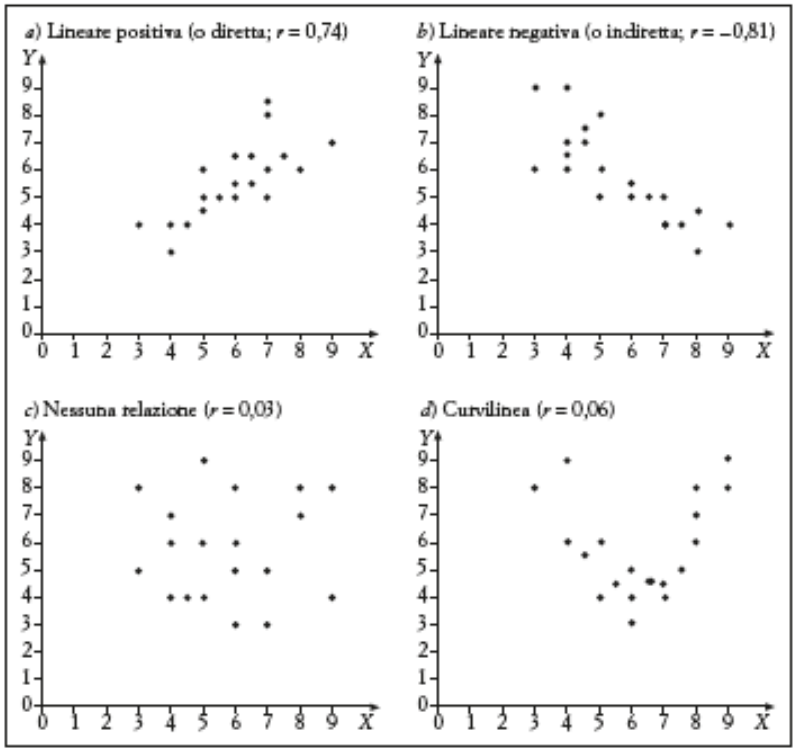

fig. 3.3. *Diagrammi di dispersione raffiguranti quattro tipi di relazioni fra due variabili.*

- Retta di regressione

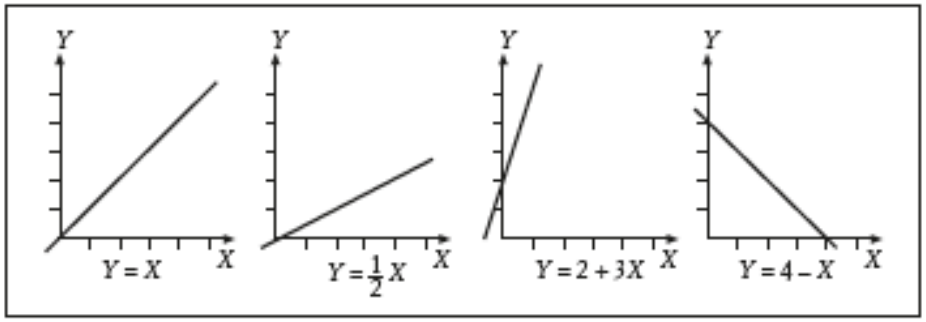

fig. 3.4. *Alcuni esempi di rette e loro equazioni.*

- Y = a + bX
- dove a è l'intercetta della retta sull'asse delle Y (cioè l'ordinata della retta quando l'ascissa è 0) e b è l'inclinazione della retta (cioè la variazione dell'ordinata quando l'ascissa varia di un'unità)

Coefficiente di correlazione

- coefficiente di correlazione di Pearson

$$r = \frac{\Sigma(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\Sigma(X_i - \overline{X})^2 \, \Sigma(Y_i - \overline{Y})^2}}$$

- r assume valore +1 in caso di relazione perfetta positiva, –1 in caso di relazione perfetta negativa, e 0 in caso di assenza di relazione
- r è un numero puro, nel senso che non risente dell'unità di misura delle due variabili

- La relazione fra due variabili X e Y può risultare dai dati, ma tuttavia può non essere dovuta a un affetto di causazione (X causa Y), in quanto può esistere una terza variabile Z che influenza entrambe, producendo una correlazione fra X e Y che non deve essere interpretata in termini di causazione fra X e Y.
- Per il rischio di questo errore, anche nel caso di analisi bivariate **è sempre opportuno introdurre nell'analisi terze variabili**, allo scopo di purificare e chiarire la relazione fra le due variabili iniziali X e Y.
- A seconda del modo di interagire della terza variabile con le prime due, la relazione fra X e Y può risultare:
  - **spuria,**
  - **Indiretta**
  - **condizionata**