

Statistica per la ricerca sociale



Roberto Albano
Silvia Testa

Carocci editore

Fondamenti e metodi per l'analisi empirica nelle scienze sociali a.a. 2013-2014
Corso di laurea in *Teorie, culture e tecniche per il Servizio sociale*
Università degli Studi di Macerata

gnificatività. La distribuzione campionaria di questo indice è asintoticamente normale con media 0 e scarto quadratico medio $\frac{1}{\sqrt{n-1}}$.

La statistica test in questo caso è data dal valore z , calcolato nel modo consueto e confrontato con la tavola della normale standardizzata.

4. Analisi della varianza

L'analisi della **varianza** o ANOVA (acronimo inglese di *ANalysis Of VAriance*) è una tecnica statistica nata nell'ambito della ricerca sperimentale per valutare l'effetto di determinati fattori, o variabili indipendenti, su una variabile dipendente. Negli anni trenta del secolo scorso Fisher e suoi collaboratori svilupparono la tecnica dell'analisi della varianza all'interno di un centro di sperimentazione agraria, allo scopo di studiare l'effetto di determinati fattori sulla produzione agricola.

L'analisi della varianza, nata in un contesto sperimentale, ha trovato ampia applicazione anche in contesti non sperimentali (molto frequenti soprattutto nella ricerca sociale), diventando oggi una delle procedure più utilizzate per valutare, generalmente, relazioni di influenza tra variabili.

L'analisi della varianza assume nomi diversi a seconda di quante sono le variabili dipendenti e indipendenti. In particolare si parla di:

- ANOVA a una via (*One-Way ANOVA*) quando si ha una sola variabile dipendente e una sola variabile indipendente;
- ANOVA fattoriale quando si ha una sola variabile dipendente, ma più di una variabile indipendente;
- MANOVA (*Multivariate ANalysis Of VAriance*) quando sia le dipendenti sia le indipendenti sono più di una.

Dato il carattere introduttivo del testo soffermeremo la nostra attenzione unicamente sul caso più elementare di analisi della varianza, quella ad una via. Inoltre limiteremo l'attenzione al caso degli **effetti fissi**, ossia la situazione in cui tutte le modalità dell'indipendente vengono introdotte nell'analisi (per la cosiddetta ANOVA a **effetti random**, in cui le modalità considerate sono un sottoinsieme casuale di quelle possibili, cfr. Iversen, 1976). Quando è opportuno ricorrere all'analisi della varianza? Dal punto di vista degli scopi di una ricerca, è la tecnica più idonea quando si desideri valutare l'esistenza di differenze tra le medie di più gruppi tra loro indipendenti, o, in modo del tutto equivalente, quando si voglia valutare all'interno di un solo campione la relazione di influenza di una variabile indipendente su una variabile dipendente, con il vincolo che la variabile indipendente assuma un numero ridotto di modalità distinte (ordinate o non).

Dal punto di vista statistico l'ambito di applicazione viene drasticamente ridotto. Una prima restrizione è dovuta al tipo di variabili coinvolte: la va-

Nata in ambito
sperimentale,
l'ANOVA si applica
oggi anche in
ricerche non
sperimentali

riabile dipendente deve essere cardinale, anzi più rigorosamente dovrebbe essere una variabile continua e illimitata (ossia una variabile a livello di scala di intervalli); la variabile indipendente può essere invece di qualsiasi tipo, ma deve essere ricondotta ad una variabile con poche categorie, eventualmente mediante un raggruppamento in classi.

Una seconda restrizione è comune a tutti i test parametrici, tra i quali rientra senz'altro l'ANOVA, ed è legata al rispetto di alcuni assunti molto stringenti. Per meglio comprendere gli assunti è opportuno dapprima familiarizzare con la simbologia tipicamente adottata in questo contesto. Iniziamo dal modo in cui possono presentarsi i dati. Un primo modo è quello tradizionale della matrice $C \times V$ (cfr. fig. 1 del cap. 1).

Con la lettera Y indichiamo la variabile dipendente, con la lettera X quella indipendente, che presenta K modalità (1, 2, ..., k , ..., K). Con i indichiamo un generico caso e con N il numero complessivo di casi (1, 2, ..., i , ..., N). Quando per ogni modalità della variabile X si presenta lo stesso numero di casi si parla di **disegno bilanciato**. In questo caso possiamo organizzare i dati nella **matrice di calcolo** di tab. 9.

TABELLA 9 Matrice di calcolo nell'ANOVA

Casi	x_1	Modalità della variabile indipendente					
			x_k		x_K		
1	y_{11}		y_{1k}		y_{1K}		
:	.		.				
i	y_{i1}		y_{ik}		y_{iK}		
:	.		.				
n_j	y_{n1}		y_{nj}		y_{nK}		

Il vettore della variabile Y di dimensione N è stato suddiviso in K vettori di dimensione n_k (dove n_k indica la numerosità del k -esimo gruppo) successivamente accostati uno all'altro. Le N osservazioni sono così state raggruppate in K gruppi, in modo che $N = \sum n_k$, e se, come nel nostro caso, il disegno è bilanciato ($n_1 = n_2 = \dots = n_K$) $N = K \cdot n$, dove n privo di pedice indica la numerosità di ciascun gruppo. I gruppi differiscono per il valore assunto dalla variabile indipendente X . Con questa organizzazione risulta più immediato poter pensare che i valori osservati per la variabile dipendente Y (le celle di questa seconda matrice) provengano da K popolazioni, ossia da tante popolazioni quante sono le modalità della variabile indipendente.

Introduzione alla statistica per la ricerca sociale

Il ricorso all'ANOVA è legittimo se sono soddisfatti alcuni assunti

- Possiamo a questo punto elencare gli assunti dell'ANOVA:
- i gruppi devono essere indipendenti;
- la variabile Y deve avere la stessa varianza nelle K popolazioni (assunto di omoschedasticità);
- la variabile Y deve distribuirsi normalmente nelle K popolazioni (assunto di normalità).

Questo nelle scienze sociali è un problema perchè di solito le variabili non sono distribuite normalmente, anzi non sono proprio cardinali

L'ANOVA può essere usata sia per testare ipotesi sia per stimare parametri

Qual è la logica dell'ANOVA? Prima di rispondere a questa domanda è necessario restringere ancora ulteriormente la nostra trattazione. L'ANOVA è una tecnica sofisticata che permette sia di condurre dei test di verifica di ipotesi (come vedremo, essa può essere considerata una estensione del test della differenza tra due medie), sia di compiere delle stime puntuali ed intervallari, risultando, in questo secondo caso, del tutto affine alla tecnica della regressione (che verrà discussa nel prossimo paragrafo).

Si potrebbe infatti sviluppare l'analisi della varianza come un particolare esempio di modello lineare; il discorso verrà qui limitato comunque all'ANOVA come test di ipotesi statistiche (per una trattazione dell'ANOVA come modello lineare rinviamo a Miceli, 2001).

Data questa delimitazione possiamo riformulare la domanda nel seguente modo: quali ipotesi vengono sottoposte a verifica e qual è il ragionamento che porta all'accettazione o al rifiuto di esse? Nell'ANOVA le ipotesi sono le seguenti:

$$H_0: \mu_1 = \mu_2 = \dots = \mu_K$$

H_1 : almeno due medie sono tra loro differenti

Procediamo a sviluppare le conseguenze che derivano dal ritenere validi gli assunti e vera l'ipotesi nulla. Se valgono gli assunti, le K popolazioni, da cui provengono i rispettivi gruppi, possono differire unicamente per il parametro μ (si ricorda che una distribuzione normale dipende unicamente dai parametri μ e σ); se l'ipotesi nulla è vera le K popolazioni coincidono per la caratteristica in esame: ogni gruppo proviene allora dalla stessa popolazione in cui la caratteristica in esame (variabile dipendente) è distribuita secondo la funzione $N(\mu, \sigma)$. Poiché ci troviamo di fronte a più gruppi appartenenti alla stessa popolazione, abbiamo due modi indipendenti di stimare la varianza della popolazione.

Secondo una prima modalità, calcoliamo le medie all'interno di ciascun gruppo (dette **medie parziali** o condizionate) e consideriamo i K valori ottenuti come determinazioni della variabile aleatoria media campionaria. Poiché la varianza della media campionaria è σ^2/n , otteniamo una stima della varianza della popolazione moltiplicando la varianza della media campionaria per n.

Una seconda modalità consiste nello stimare σ^2 con la media delle varianze

(corrette) calcolate all'interno di ciascun gruppo. La stima ottenuta nel secondo modo è in realtà indipendente dalla veridicità dell'ipotesi nulla, mentre quella ottenuta con il primo è una stima non affetta da errore sistematico (distorsione) soltanto se l'ipotesi nulla è vera (soltanto in questo caso le K medie parziali sono realizzazioni campionarie della media campionaria).

Un modo quindi del tutto naturale di controllare la plausibilità dell'ipotesi nulla è quello di costruire il rapporto tra le due stime indipendenti di σ^2 e osservare quanto questo rapporto si discosta da 1, ossia dalla condizione di uguaglianza.

Sappiamo che la statistica costruita come rapporto tra due varianze corrette si distribuisce come la v.a. F di Fisher-Snedecor, ed è proprio in base alla tavola statistica F che andremo a valutare la significatività dello scostamento da 1 del valore restituito da tale rapporto. Se lo scostamento non è significativo accettiamo l'ipotesi nulla, in caso contrario la rifiutiamo.

Riassumendo, possiamo dire che nell'ANOVA per controllare l'effetto dei diversi livelli della variabile indipendente, ossia per testare la differenza tra le medie dei gruppi, sfruttiamo la possibilità di separare due fonti di variabilità nella variabile dipendente, una attribuibile alla variabile indipendente – presente nella **varianza tra i gruppi** (varianza *between*), ma non nella **varianza entro i gruppi** (varianza *within*) – e una attribuibile alla fluttuazione casuale, presente sia nella varianza tra i gruppi sia in quella entro i gruppi; dal confronto delle due varianze possiamo decidere se la variabile indipendente influenza quella dipendente.

Passiamo ora alle formule che ci permetteranno di calcolare le due varianze e ricavarne il rapporto. Con $\bar{y}_k$ indichiamo la media del gruppo k -esimo:

$$\bar{y}_k = \frac{1}{n} \sum_{i=1}^n y_{ij}$$

Due fonti di
variabilità della v .
dipendente: v .
indipendente e
errore casuale

La presenza del pedice k indica che la media è stata condotta su tutti i casi del k -esimo gruppo. Con $\bar{y}$ indichiamo la media generale (cioè calcolata su tutti gli N casi osservati, indipendentemente dalla loro distinzione in gruppi):

$$\bar{y} = \frac{1}{K} \sum_{k=1}^K \bar{y}_k$$

Prima di calcolare le varianze soffermiamoci sui loro numeratori, ossia sulle devianze.

Introduzione alla statistica per la ricerca sociale

Possiamo in primo luogo calcolare una devianza dei valori di Y ignorando la distinzione nei K gruppi; chiameremo questa quantità **devianza totale**:

$$\text{Dev}_T = \sum_{i=1}^n \sum_{k=1}^K (y_{ik} - \bar{y})^2$$

Essa esprime la variabilità delle osservazioni attorno alla loro media generale.

Allo stesso modo possiamo calcolare la variabilità delle medie di gruppo attorno alla loro media, che è ancora la media generale; chiameremo questa quantità **devianza tra le medie**:

$$\text{Dev}_M = \sum_{k=1}^K (\bar{y}_k - \bar{y})^2$$

Abbiamo ancora una possibilità, valutare la variabilità delle osservazioni attorno alla loro media di gruppo; chiameremo questa quantità **devianza parziale tra le osservazioni**:

$$\text{Dev}_P = \sum_{i=1}^n \sum_{k=1}^K (\bar{y}_{ik} - \bar{y}_k)^2$$

che esprime la variabilità presente in ciascun gruppo.

Possiamo sfruttare queste due ultime quantità per giungere a due stime indipendenti della varianza della popolazione da cui si suppone provengano i K gruppi.

Calcolo della
varianza tra i gruppi

Per trasformare la Dev_M in una varianza corretta non bisogna fare altro che dividerla per il numero dei gradi di libertà (pari a $K-1$), ottenendo in questo modo una stima della varianza della distribuzione della media campionaria; per passare dalla varianza della distribuzione della media campionaria a quella della popolazione è sufficiente moltiplicare la prima per il numero di osservazioni (n). Effettuando entrambe le operazioni otteniamo la prima stima della varianza della popolazione che chiamiamo varianza tra i gruppi:

$$\text{VAR}_B = \frac{n \sum_{k=1}^K (\bar{y}_k - \bar{y})^2}{K - 1}$$

Calcolo della
varianza entro i
gruppi

Attraverso la devianza parziale tra le osservazioni perveniamo invece alla seconda stima della varianza della popolazione. Dapprima, nell'ambito di ciascun gruppo, trasformiamo la devianza in varianza dividendo per i gradi

di libertà $(n - 1)$ e in seguito calcoliamo una media di queste stime locali, ottenendo la varianza entro i gruppi:

$$\begin{aligned} \text{VAR}_w &= \frac{1}{K} \sum_{k=1}^K \frac{\sum_{i=1}^n (y_{ik} - \bar{y}_k)^2}{n - 1} = \frac{\sum_{k=1}^K \sum_{i=1}^n (y_{ik} - \bar{y}_k)^2}{K(n - 1)} = \\ &= \frac{\sum_{k=1}^K \sum_{i=1}^n (y_{ik} - \bar{y}_k)^2}{K(n - K)} = \frac{N - K}{N - K} \end{aligned}$$

Se l'ipotesi nulla è vera le due stime hanno lo stesso valore atteso, pari alla varianza della popolazione; se l'ipotesi nulla è falsa il valore atteso della varianza tra i gruppi sarà la somma della varianza della popolazione e della varianza delle medie delle popolazioni a cui i diversi gruppi appartengono. Si dice allora che la varianza tra i gruppi è affetta da errore sistematico (è uno stimatore distorto della varianza della popolazione). Di conseguenza, il rapporto tra la varianza tra i gruppi e quella entro i gruppi ha come valore atteso 1 se l'ipotesi nulla è vera.

Siamo così giunti alla statistica test:

$$F = \frac{\text{VAR}_B}{\text{VAR}_w}$$

la cui distribuzione campionaria è quella della v.a. F di Fisher-Snedecor, con $(K - 1)$ e $(N - K)$ gradi di libertà.

Tenendo conto del livello di significatività α , fissato a priori, H_0 non è accettata quando il valore di F cade nella zona di rifiuto, ossia è maggiore di F_α , dove F_α è il valore ottenuto dalla tavola della distribuzione F per $K - 1$ e $N - K$ gradi di libertà.

È buona norma affiancare ad una eventuale significatività statistica una misura della strettezza della relazione. Poco importa infatti sapere ad esempio che la differenza nella capacità di rievocare delle informazioni è significativamente influenzata dal tipo di elaborazione dell'informazione, se questa influenza spiega una quota ridottissima della variabilità esistente nella capacità di rievocazione. Occorre inoltre tenere conto che una delle caratteristiche indesiderate dei test statistici in generale consiste nel fatto di portare praticamente sempre al rifiuto dell'ipotesi nulla quando il numero delle osservazioni è elevato.

Per valutare la strettezza della relazione possiamo utilizzare l'operatore E^2 incontrato nel cap. 3, trattando degli operatori di determinazione:

$$E^2 = \frac{\text{Dev}_B}{\text{Dev}_w}$$

Significatività
statistica e misura
della strettezza di
una relazione tra
variabili

Introduzione alla statistica per la ricerca sociale

Il teorema di scomposizione della devianza non si applica alle varianze corrette

Nel contesto inferenziale, quando si ricorre alla formula corretta per il calcolo delle varianze, non è più possibile parlare in modo equivalente di quota di devianza della variabile dipendente riprodotta dalla variabile indipendente o di quota di varianza riprodotta, in quanto i gradi di libertà del numeratore e del denominatore sono differenti; il teorema di scomposizione della devianza, presentato nel cap. 3, non può più essere esteso alle varianze corrette. Per ottenere una misura della strettezza della relazione, nei termini di quota di varianza (corretta per i gradi di libertà) riprodotta possiamo ricorrere all'operatore **eta quadro**:

$$\eta^2 = 1 - \frac{VAR_W}{VAR_T}$$

η^2 ed E^2

Questo operatore non è altro che la controparte inferenziale del coefficiente di determinazione E^2 ; a differenza di quest'ultimo, consente di tenere conto dell'ampiezza del campione e del numero di modalità della variabile indipendente. η^2 assume come valore massimo 1, nel caso la variabile indipendente sia in grado di riprodurre perfettamente la variabile dipendente; se le variabili sono indipendenti in media il suo valore sarà un numero non troppo distante da zero.

Per facilitare la comprensione dell'utilizzo dell'analisi della varianza viene di seguito riportato un esempio, ispirato a una ricerca sulla memoria realmente condotta da Eysenck (come risulta dal suo saggio *Age Differences in Incidental Learning*, in "Developmental Psychology", 10, 1974, pp. 936-41). Eysenck era interessato alla verifica empirica del modello teorico di Craik e Lockhart, secondo il quale per la capacità di rievocare e riconoscere l'informazione è di fondamentale importanza il livello di elaborazione dell'informazione stessa. A questo scopo suddivise casualmente un gruppo di 50 studenti in 5 gruppi, ognuno assegnato ad una condizione sperimentale differente. Data una lista di 20 parole, le 5 condizioni consistevano nei seguenti compiti:

- contare il numero di lettere contenute in ogni parola;
- cercare una rima per ogni parola;
- cercare un aggettivo che cambiasse il senso della parola;
- creare una immagine mentale;
- memorizzare le parole per rievocarle in seguito.

Dopo la presentazione della lista veniva richiesto ai 5 soggetti di rievocare quante più parole possibile. Delle cinque condizioni sperimentali in una soltanto l'apprendimento era intenzionale (l'ultima); inoltre, le altre quattro condizioni differivano tra loro per grado di profondità di elaborazione, richiedendo il contare il numero di lettere un livello di elaborazione del-

l'informazione più superficiale degli altri e la creazione di una immagine visiva il livello più profondo.

La variabile indipendente è quindi il “tipo di elaborazione dell'informazione”, una variabile categoriale a cinque modalità o livelli di trattamento; la variabile dipendente è il “numero di parole rievocate”, un conteggio che può assumere valori tra 0 e 20. Per non appesantire i calcoli, immaginiamo che i livelli siano soltanto 3 e che ogni gruppo sia formato da 3 studenti.

L'ipotesi nulla formulata dallo psicologo è $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ e l'ipotesi alternativa è H_1 ; almeno due medie sono differenti. L'ipotesi H_1 asserisce che diversi livelli di elaborazione dell'informazione producono differenze nella rievocazione, differenze che si riflettono in un differente numero medio di parole rievocate nei diversi gruppi; con l'ipotesi nulla si afferma invece che il tipo di elaborazione non ha alcun effetto sul numero di parole rievocate. I risultati ottenuti sono riassunti nella tab. 10.

TABELLA 10 Matrice di calcolo: esempio numerico

	Modalità della variabile indipendente			Totale
	L1	L2	L3	
	6	8	4	
Osservazioni	7	9	6	
	5	7	5	
Somme	18	24	15	57
Medie	6	8	5	6,33
N. osservaz.	3	3	3	9

I calcoli per verificare H_0 sono i seguenti:

$$\begin{aligned} \text{Dev}_T &= (6 - 6,33)^2 + (7 - 6,33)^2 + \dots + (5 - 6,33)^2 = 20 \\ \text{Dev}_B &= 3[(6 - 6,33)^2 + (8 - 6,33)^2 + (5 - 6,33)^2] = 14 \\ \text{Dev}_W &= [(6 - 6)^2 + (7 - 6)^2 + (5 - 6)^2] + [(8 - 8)^2 + (9 - 8)^2 + (7 - 8)^2] + \\ &\quad + [(4 - 5)^2 + (6 - 5)^2 + (5 - 5)^2] = 6 \end{aligned}$$

Possiamo osservare numericamente che, per le devianze, vale il teorema di scomposizione:

$$\text{Dev}_T = \text{Dev}_B + \text{Dev}_W = 14 + 6 = 20$$

Introduzione alla statistica per la ricerca sociale

Passiamo ora al calcolo dei gradi di libertà:

$gdl_T = N - 1 = 8$
 $gdl_B = K - 1 = 2$
 $gdl_W = (n - 1)K = 6$

e a quello delle varianze:

$VAR_T = 20/8 = 2,5$
 $VAR_B = 14/2 = 7$
 $VAR_W = 6/6 = 1$

Infine, calcoliamo il valore della statistica test:

$F = \frac{VAR_B}{VAR_W} = 7$

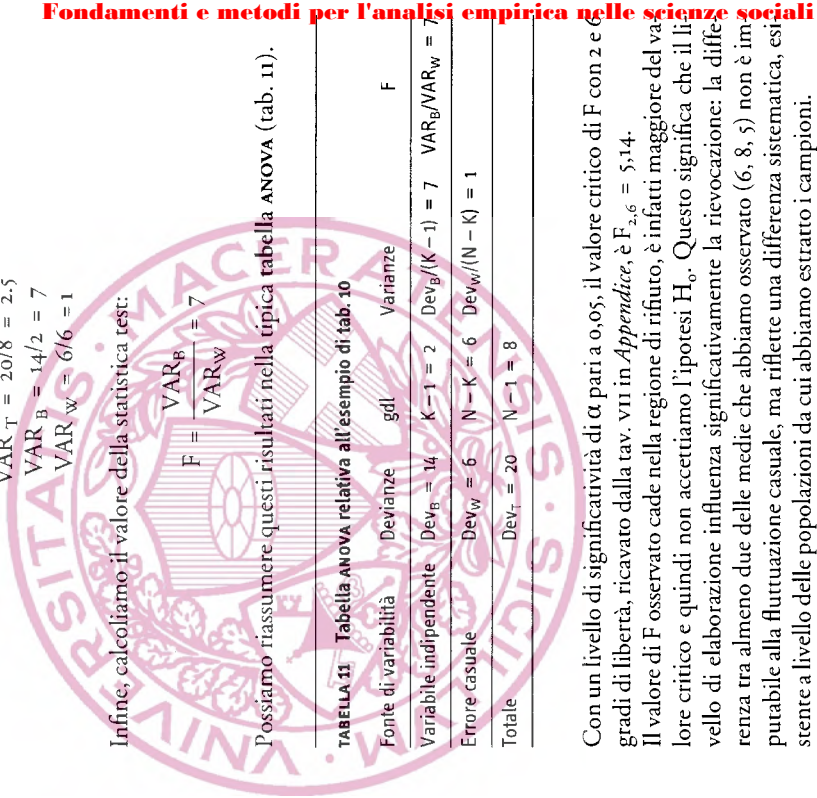
Possiamo riassumere questi risultati nella tipica **tabella ANOVA** (tab. 11).

TABELLA 11 Tabella ANOVA relativa all'esempio di tab. 10				
Fonte di variabilità	Devianze	gdl	Varianze	F
Variabile indipendente	$Dev_B = 14$	$K - 1 = 2$	$Dev_B/(K - 1) = 7$	$VAR_B/VAR_W = 7$
Errore casuale	$Dev_W = 6$	$N - K = 6$	$Dev_W/(N - K) = 1$	
Totale	$Dev_T = 20$	$N - 1 = 8$		

Con un livello di significatività di α pari a 0,05, il valore critico di F con 2 e 6 gradi di libertà, ricavato dalla tav. vii in *Appendice*, è $F_{2,6} = 5,14$. Il valore di F osservato cade nella regione di rifiuto, è infatti maggiore del valore critico e quindi non accettiamo l'ipotesi H_0 . Questo significa che il livello di elaborazione influenza significativamente la rievocazione: la differenza tra almeno due delle medie che abbiamo osservato (6, 8, 5) non è imputabile alla fluttuazione casuale, ma riflette una differenza sistematica, esistente a livello delle popolazioni da cui abbiamo estratto i campioni. Possiamo chiederci se oltre ad essere statisticamente significativa l'influenza della variabile indipendente sia anche forte, ricorrendo alla già menzionata misura η^2 :

$$\eta^2 = 1 - \frac{VAR_W}{VAR_T} = 1 - \frac{1}{2,5} = 1 - 0,4 = 0,6$$

In base a questa statistica la relazione sembra essere forte.



L'esempio appena visto, ci permette di fare alcune ulteriori considerazioni teoriche. In primo luogo la variabile dipendente non è continua, essendo un conteggio; ciò può rendere decisamente implausibile l'assunto di normalità (il rischio è tanto maggiore quanto minore è il numero di modalità che il conteggio effettivamente assume) e può portare ad una violazione anche dell'assunto di omoschedasticità; è sufficiente ricordare che nelle variabili discrete media e varianza non sono indipendenti (ad esempio nella binomiale $E(Y) = np$ e $VAR(Y) = npq$) e quindi differenze nelle medie si ripercuoteranno anche sulle varianze. Soltanto se l'ampiezza del campione è grande si può utilizzare con tranquillità anche la scala assoluta, potendo sfruttare l'approssimazione alla distribuzione normale anche delle distribuzioni discrete (per il teorema del limite centrale). Per ovviare alla dipendenza tra medie e varianze e ad altre ragioni di scostamento delle distribuzioni dalla normale si può ricorrere ad una trasformazione della variabile dipendente stessa, operazione che deve comunque essere condotta con cautela, non essendo innocua rispetto ai risultati dell'ANOVA (per una trattazione del problema della trasformazione della variabile dipendente si può consultare Ercolani, Areni, Mannetti, 1990, pp. 119-22; Howell, 1992, pp. 307-16; per i rischi connessi a questo genere di manipolazione cfr. McBurney, 1983, trad. it. pp. 244-8). Al di là della natura continua o discreta della variabile è comunque necessario che essa abbia una distribuzione il più simmetrica possibile e soprattutto che per i diversi gruppi non sia asimmetrica in direzione opposta.

In realtà è invalso un uso abbastanza disinvolto di questa tecnica, giustificato solo in parte. È vero infatti che l'ANOVA è una tecnica robusta, ossia in grado di produrre risultati validi anche in presenza di violazioni degli assunti, ma questo non significa che questi ultimi si possano tranquillamente ignorare. È buona norma domandarsi quanto i nostri dati sono adeguati per una analisi della varianza, ispezionando graficamente la distribuzione dei dati, testando l'omogeneità delle varianze (a tal fine si utilizza un test detto di Levene). Come suggerisce Howell (1992), se le popolazioni possono essere considerate simili nella forma e la varianza più grande non supera di 4 volte quella più piccola, l'analisi può risultare valida. Se si sospetta l'inapplicabilità dell'ANOVA è meglio ricorrere ad un analogo test non parametrico (per esempio si ricorre al *test di Kruskal-Wallis* quando la variabile dipendente è ordinale: cfr. Blalock, 1960, trad. it. pp. 461-7).

Quando le modalità della variabile indipendente sono due soltanto, l'analisi della varianza produce risultati equivalenti a quelli del test della differenza tra due medie per campioni indipendenti. Infatti, in questo caso tra le due statistiche test esiste la seguente relazione:

$$F = t^2$$

L'ANOVA è una tecnica robusta, ma gli assunti non vanno ignorati

Esistono alternative non parametriche all'ANOVA

Ci si potrebbe chiedere allora perché non utilizzare il test della differenza tra le medie anche quando i gruppi sono più di due, semplicemente replicandolo più volte. La risposta è connessa alla probabilità di commettere un errore di 1° tipo: se in una ricerca si ricorre più volte al test della differenza tra le medie fissando ad esempio un livello di significatività dello 0,05, la probabilità di rifiutare l'ipotesi nulla quando essa è vera, considerando non il singolo confronto ma tutti quelli eseguiti, non è più pari a 0,05 bensì superiore; di quanto cresce, dipende dal numero di test eseguiti.

Detto in termini più discorsivi: posto alfa pari a 0,05, condotti 100 test, e posto che per tutti i confronti l'ipotesi nulla sia vera, dobbiamo aspettarci di rifiutare erroneamente l'ipotesi nulla in 5 casi su 100; in 5 casi le medie risulteranno significativamente diverse tra di loro, per puro effetto del caso. Soffermiamoci non sul numero di errori che possiamo compiere, ma sulla probabilità di compiere almeno un errore del 1° tipo in un insieme di test; probabilità che indicheremo con F_w (*family wise error*). Questa è data dalla seguente espressione:

$$F_w = 1 - (1 - \alpha)^c$$

dove c è il numero di confronti eseguiti. Si può così notare come questa probabilità si discosti notevolmente dal valore prescelto per ogni singolo confronto quando il numero dei gruppi cresce, e coincide con α quando $c = 1$ (ossia nel caso vengano confrontati soltanto due gruppi). Con l'ANOVA effettuiamo invece un solo test e la sua probabilità di errore del primo tipo coincide con il livello di significatività prescelto. A questo indubio vantaggio statistico si contrappone uno svantaggio legato allo scopo del test. Al ricercatore molto spesso non è sufficiente sapere che tra i gruppi esistono delle differenze significative, ma preme sapere quali sono in specifico i gruppi che differiscono. Per ottenere questa informazione le strade possibili sono due:

- soffermare, fin dalla progettazione dell'esperimento, l'attenzione su di un numero ridotto di confronti, eseguendo pochi test di differenza tra le medie (situazione detta "dei confronti a priori");
- una volta ottenuto un test F significativo, eseguire tutti i confronti possibili, ma con delle procedure statistiche, definite **test post hoc** in grado di controllare la probabilità dell'errore di 1° tipo.

5. Regressione lineare

Quando studiamo le relazioni tra variabili cardinali, oltre alla forza e alla significatività dell'associazione, siamo interessati alla sua forma,

possibilmente riconducibile ad una relazione funzionale analiticamente nota.

Queste tre informazioni che possono caratterizzare il rapporto tra variabili, considerate nel loro complesso, costituiscono un **modello statistico**. La **regressione lineare** appartiene alla famiglia dei modelli lineari, la cui caratteristica è quella di esprimere una variabile dipendente Y , misurata a livello di scala cardinale, come funzione lineare dei seguenti termini espliciti:

- una o più variabili indipendenti X_K , anch'esse cardinali e osservate per almeno due valori distinti (possono quindi anche essere delle variabili *dummy*: cfr. Hardy, 1993); tali variabili sono anche dette variabili esplicite o regressori;

- una componente erratica, che sintetizza l'effetto di tutte le fonti di variabilità non previste nel modello e quello dell'errore di misurazione della variabile Y .

La regressione è una tecnica statistica utile per valutare se, e in che misura, si possa parlare dell'esistenza di un rapporto di dipendenza tra variabili; tuttavia, il fatto di aver verificato che tale dipendenza esiste non comporta necessariamente che tra le variabili esista un rapporto causale. Il rapporto di causa-effetto si pone su un piano semantico dell'interpretazione dei dati, in quanto discende da un ragionamento logico-teoretico, mentre la dipendenza tra variabili è stabilita su un piano sintattico (cosa che peraltro vale per ogni altra tecnica in cui si distingue tra variabili dipendenti e variabili indipendenti). Fatta questa distinzione, si può però anche osservare che tra l'analisi della causalità e le tecniche come la regressione esiste un'affinità elettiva.

La regressione lineare è una tecnica di analisi molto utilizzata nelle scienze sociali ed economiche; inoltre, molte altre tecniche di analisi hanno come nucleo la struttura matematica della regressione o una sua elaborazione (per esempio quelle citate in apertura di capitolo). In forma matematica la regressione lineare si esprime mediante la seguente equazione:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + \varepsilon_i$$

o anche, in forma più compatta:

$$Y_i = \alpha + \sum_{k=1}^K \beta_k X_{ki} + \varepsilon_i$$

dove K rappresenta il numero di variabili indipendenti incluse nel modello.

Se la variabile indipendente è una sola, si parla di **regressione lineare semplice**, altrimenti di **regressione lineare multipla**.

Definizione di
modello statistico

Il modello di
regressione lineare è
alla base di molte
altre tecniche di
analisi dei dati

Il grado complessivo del polinomio che compare a destra dell'equazione è pari a uno; quando è maggiore di uno, come nel caso seguente:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} X_{2i} + \beta_4 X_{2i}^2 + \dots + \varepsilon_i$$

si entra nel campo della non linearità. Per analizzare le relazioni non lineari esistono diverse tecniche, alcune delle quali sono delle semplici estensioni del modello di regressione lineare multipla; tuttavia, dato il carattere introduttivo di questo manuale, tralasceremo queste complicazioni (per un'introduzione ai modelli non lineari cfr. Bohrnstedt, Knoke, 1994).

Per semplicità, considereremo ora il caso della regressione lineare semplice (quindi con una sola variabile indipendente), trascurando in un primo momento la componente erratica; ciò significa assumere che la relazione tra le due variabili sia completamente deterministica, cioè che il valore della Y si possa predire esattamente una volta noto il valore della X e quello dei parametri dell'equazione. Scriviamo pertanto:

$$Y_i = \alpha + \beta X_i$$

Significato
geometrico dei
parametri del
modello di
regressione

Come si vede, questa è la formula generale di una retta avente come intercetta il valore α e come coefficiente d'inclinazione il valore β .

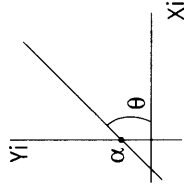
Il significato di β può essere facilmente compreso con una serie di esempi:

- se $\beta = 0,5$, un aumento di una unità della variabile X comporta una variazione positiva nella variabile Y di mezza unità;
- se $\beta = 1$, a una variazione unitaria della variabile indipendente corrisponde una variazione unitaria positiva della dipendente (come nel caso della retta rappresentata nella figura 1);
- se $\beta = 2$, a una variazione unitaria della variabile X corrisponde un incremento di due unità della Y ;
- se $\beta = -0,5$, l'entità della variazione è in modulo la stessa di quella del primo caso considerato, ma la variazione della variabile Y è di segno opposto.

Il termine β è detto anche coefficiente angolare, in quanto è proprio dal valore che esso assume che dipende l'angolo che la retta forma con l'asse delle ascisse. Un $\beta = 1$ per esempio corrisponde a una inclinazione di 45° rispetto all'asse delle ascisse; più in generale: $\beta = \tan \theta$, dove θ è l'angolo formato dalla retta con l'asse delle ascisse.

Il termine α rappresenta invece il valore che Y assume quando X è uguale a 0; ovviamente è possibile che α sia uguale a zero, nel qual caso avremo una retta che passa per l'origine.

FIGURA 1 Retta di regressione di Y su X: rappresentazione geometrica dell'intercetta α e dell'angolo θ



Con l'equazione di regressione, una volta noti i parametri α e β è pertanto possibile predire il valore assunto dalla variabile dipendente Y dato un certo valore di X. Si noti che la possibilità di predizione è frutto di una certa relazione matematica esistente tra variabili, cosa ben diversa dall'affermare che X è una causa di Y, affermazione attinente al piano teorico e non a quello empirico.

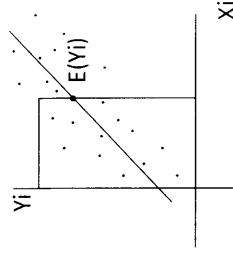
Quando si passa dallo studio analitico della retta all'esame di una distribuzione empirica, per quanto quest'ultima si approssimi alla linearità non ci si può aspettare che i valori della variabile Y siano esattamente allineati sulla retta. Anche se arriviamo a specificare un'equazione di regressione particolarmente complessa, contenente al limite tutte le variabili X rilevanti per spiegare un fenomeno Y, ci sarà sempre un elemento di fluttuazione accidentale (per esempio un errore di misurazione).

Sotto particolari assunti (di cui diremo fra breve) possiamo però attenderci una situazione del seguente tipo:

$$E(Y_i) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki}$$

Si dice allora che la retta di regressione interpola al meglio una nube di punti (fig. 2).

FIGURA 2 Nube di punti e retta interpolante



Le coordinate tracciate nel grafico servono a evidenziare il fatto che dato un certo valore x_i , il corrispondente valore y_i non cade esattamente sulla

Introduzione alla statistica per la ricerca sociale

retta (punto che possiamo chiamare y_r), ma in sua prossimità. La differenza $\varepsilon_i = Y_i - E(Y_i)$ è l'errore stocastico.

Matematicamente la retta che interpola al meglio i nostri dati si individua tipicamente con il metodo di stima dei minimi quadrati ordinari. Questo è particolarmente interessante da un punto di vista statistico in quanto, sotto determinati assunti, esso fornisce per l'intercetta e i coefficienti angolari i migliori stimatori lineari non distorti:

- lineari, perché sono funzioni lineari delle osservazioni;
- non distorti, perché il valore atteso coincide con il parametro da stimare;
- migliori, perché sono più efficienti di ogni altro stimatore lineare corretto.

Gli assunti minimi necessari perché in una regressione semplice vengano prodotti stimatori di α e di β con tali caratteristiche sono i seguenti:

- $E(\varepsilon_i) = 0$, cioè che il valore atteso di ε sia uguale a zero $\forall i$;
- $\text{Cov}(X, \varepsilon_i) = 0$, cioè il residuo ε deve avere una distribuzione indipendente rispetto alla variabile indipendente X (o alle variabili indipendenti nel caso di regressione multipla); questo assunto non è necessario se la variabile indipendente è misurata senza errore stocastico (caso per la verità poco frequente nella ricerca sociale e psicologica);
- $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0, \forall i, j$;
- $\text{Var}(\varepsilon_i) = \text{Var}(\varepsilon_j) = \dots \dots \text{Var}(\varepsilon_n)$; si tratta del requisito cosiddetto di omoschedasticità.

Un ulteriore assunto sull'errore è necessario per poter effettuare dei test di significatività sulle stime di α e β :

- la distribuzione dell'errore ε deve essere normale.

Con il metodo dei minimi quadrati ordinari otteniamo una retta che minimizza la somma dei quadrati dei residui di regressione (perpendicolari all'asse delle ascisse) tra i valori empirici Y_i e i valori teorici $\hat{Y}_i$, ossia quei valori che cadono esattamente sulla retta di regressione. In termini formali:

$$\min \sum_{i=1}^n \varepsilon_i^2 \equiv \min \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Si noti che parliamo di residui (o scarti) e non di errori, e li indichiamo con la lettera latina e invece che con la lettera greca ε , in quanto questi ultimi sono entità inosservabili, mentre i residui sono entità osservate.

Trovare la retta di regressione, significa in sostanza calcolare esattamente (a meno di errori di misurazione) i due parametri α, β quando disponiamo di dati di censimento; oppure stimarli mediante due stimatori campionari a e

Il metodo dei minimi quadrati è impiegato per individuare la retta di regressione

Assunti necessari per applicare il modello di regressione semplice

Un assunto supplementare è richiesto per effettuare test di significatività

on se, come accade più spesso, stiamo lavorando su un campione. I passaggi matematici sono i seguenti:

$$\begin{aligned} Y_i &= a + bX_i + e_i \\ e_i &= Y_i - a - bX_i \\ \sum e_i^2 &\equiv \sum (Y_i - a - bX_i)^2 \end{aligned}$$

Occorre minimizzare l'ultima espressione, derivandola rispetto ad a e rispetto a b :

$$\frac{\delta}{\delta a} \sum (Y_i - a - bX_i)^2 = \sum 2(-1)(Y_i - a - bX_i) = 0$$

$$\sum (Y_i - a - bX_i) = 0$$

$$\sum Y_i - na - b \sum X_i = 0$$

$$a = \frac{\sum Y_i + b \sum X_i}{n}$$

$$a = \bar{Y} - b\bar{X}$$

$$\frac{\delta}{\delta b} \sum (Y_i - a - bX_i)^2 = \sum 2(-X_i)(Y_i - a - bX_i) = 0$$

$$\sum X_i(Y_i - a - bX_i) = 0$$

$$\sum X_i Y_i - a \sum X_i - b \sum X_i^2 = 0$$

$$\sum X_i Y_i = a \sum X_i + b \sum X_i^2$$

$$\sum X_i Y_i = (\bar{Y} - b\bar{X})n\bar{X} + b \sum X_i^2$$

$$\sum X_i Y_i = n\bar{Y}\bar{X} - b\bar{X}n\bar{X} + b \sum X_i^2$$

$$b \sum X_i^2 - bn\bar{X}^2 = \sum X_i Y_i - n\bar{Y}\bar{X}$$

$$b = \frac{\sum X_i Y_i - n\bar{Y}\bar{X}}{\sum X_i^2 - n\bar{X}^2} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

A titolo di esempio, consideriamo i seguenti dati rilevati nelle nove aree geografiche nordamericane del censimento del 1930: percentuale di neri sul totale dei residenti (dai dieci anni in su), che consideriamo come variabile indipendente, e percentuale di analfabeti, che consideriamo come dipendente.

Introduzione alla statistica per la ricerca sociale

TABELLA 12 Esempio numerico di applicazione della regressione semplice

	X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$
New England	1,1	3,7	-8,6	-1,0	73,4	8,4
Middle Atlantic	4,0	3,5	-5,7	-1,2	32,1	6,7
East North Central	3,7	2,1	-6,0	-2,6	35,6	15,4
West North Central	2,6	1,4	-7,1	-3,3	49,9	23,2
South Atlantic	27,6	8,3	17,9	3,6	321,6	65,0
East South Central	27,2	9,6	17,5	4,9	307,4	86,3
West South Central	18,8	7,2	9,1	2,5	83,4	23,0
Mountain	0,9	4,2	-8,8	-0,5	76,9	4,2
Pacific	1,1	2,1	-8,6	-2,6	73,4	22,1
	$\bar{X} = 9,7$	$\bar{Y} = 4,7$				
					$\Sigma (X_i - \bar{X})^2 = 1053,7$	$\Sigma (X_i - \bar{X})(Y_i - \bar{Y}) = 254,2$

Fonte: censimento USA 1930.

L'esame del diagramma di dispersione è utile per verificare la linearità

Determiniamo la retta di regressione, dopo avere accertato dal diagramma di dispersione che la relazione tra le variabili sia di tipo lineare. Ricaviamo ora i valori dei parametri della retta applicando le formule generali ai nostri dati:

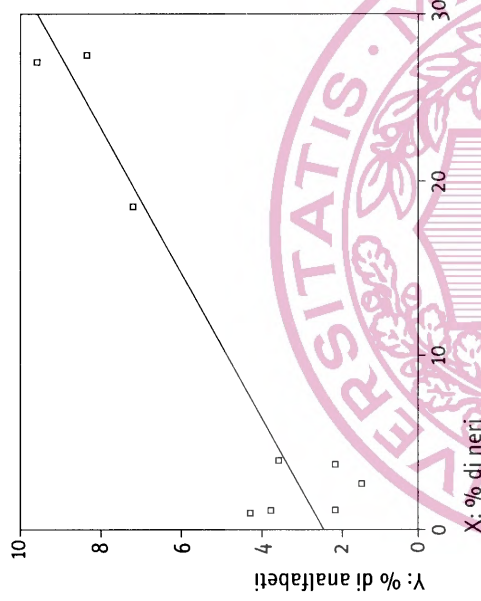
$$b = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{254,2}{1053,7} = 0,24$$
$$a = \bar{Y} - b\bar{X} = 4,7 - 0,24 \cdot 9,7 = 2,34$$

Pertanto l'equazione della retta è:

$$Y = 2,34 + 0,24 \cdot X$$

L'interpretazione da un punto di vista formale non presenta particolari difficoltà: a partire da uno "zoccolo" medio del 2,34% di analfabetismo, spiegato da altri fattori, ogni incremento unitario nella variabile dipendente produce un aumento di 0,24 punti nella percentuale di analfabeti. Ovviamente molto più complessa è l'interpretazione sostantiva di questo risultato: si potrebbe pensare che ciò sia spiegabile con il fatto che in quell'epoca i neri erano più analfabeti in media dei bianchi; ma si può pensare anche a spiegazioni alternative: per esempio che i neri vivessero in maggior misura in aree più povere, dove si addensava di più anche l'analfabetismo (di bian-

FIGURA 3 Diagramma di dispersione e retta interpolante relativi ai dati della tab. 12



chi e neri). Come si evince da questo semplice esempio, occorre evitare di confondere la “spiegazione” statistica con la spiegazione delle cause di un fenomeno: quest’ultima richiede una serie di argomentazioni più ampie, in cui certamente anche il dato statistico può giocare un ruolo rilevante.

Ma ritorniamo agli aspetti formali. Il coefficiente b dipende dalla scala di misurazione di X e Y , come si può facilmente verificare (si provi a ripetere i calcoli dell’esempio precedente cambiando scala alla variabile X , per esempio usando le frequenze relative invece delle percentuali). Per ottenere un coefficiente di regressione *a-dimensionale* (*scale-free* nel linguaggio anglosassone) si ricorre alla standardizzazione delle variabili, ottenendo quello che viene solitamente indicato come **coefficiente beta di regressione** o *coefficiente* di regressione standardizzato (da non confondersi con il parametro β , ignoto; il coefficiente di cui stiamo parlando è osservato).

Esso non è nient’altro che il coefficiente b ponderato con il rapporto tra le deviazioni standard delle variabili X e Y :

$$\text{beta}_x = b_x \cdot \frac{s_x}{s_y}$$

Si può peraltro notare che essendo b dato dal rapporto tra $\text{COV}(X, Y)$ e $\text{VAR}(X)$, nel caso della regressione semplice beta_x non è altro che il rapporto:

$$\frac{\text{COV}(Z_X, Z_Y)}{\text{VAR}(Z_X)} = \text{COV}(Z_X, Z_Y) = r_{xy}$$

ossia il coefficiente di correlazione lineare.

Introduzione alla statistica per la ricerca sociale

...tà dei
coefficienti beta
nella regressione
multipla

I coefficienti beta tornano utili nella regressione multipla; servono infatti a valutare il peso relativo che le singole variabili indipendenti hanno nello spiegare la variabilità della dipendente, se le prime sono misurate con unità di misura diverse. Se l'analisi ha come obiettivo la generalizzazione da campione a popolazione, il che richiede che il campione di dati su cui stiamo operando sia di tipo probabilistico, prima di passare a qualsiasi interpretazione di carattere sostantivo circa l'impatto dei regressori sulla variabile dipendente è necessario considerarne la significatività statistica.

Dobbiamo in altri termini sottoporre a test l'ipotesi nulla che i coefficienti di regressione siano significativamente diversi da zero. Ovviamente, ancora una volta, la scelta dell'ipotesi alternativa dipende dalla conoscenza che abbiamo del fenomeno; pertanto, se non abbiamo un'idea a priori su quale sia il segno dell'impatto del regressore sulla variabile dipendente, si opterà per un test bidirezionale, in caso contrario sarà opportuno specificare un'ipotesi unidirezionale.

Se si sta considerando un modello di regressione lineare semplice, il test sarà lo stesso già visto a suo tempo per valutare la significatività di una correlazione lineare (cfr. il par. 3.8): ciò in virtù del fatto che, come si è dimostrato poc'anzi, il coefficiente β_{1x} equivale al coefficiente di correlazione lineare; se invece si è alle prese con un modello di regressione multipla si ricorre a un test molto simile a quello esaminato nel par. 6.4 del cap. 6, cioè il test della media di una popolazione (per una illustrazione del test di significatività dei coefficienti di una regressione multipla rimandiamo a Bohrnstedt, Knoke, 1994, trad. it. pp. 192-5).

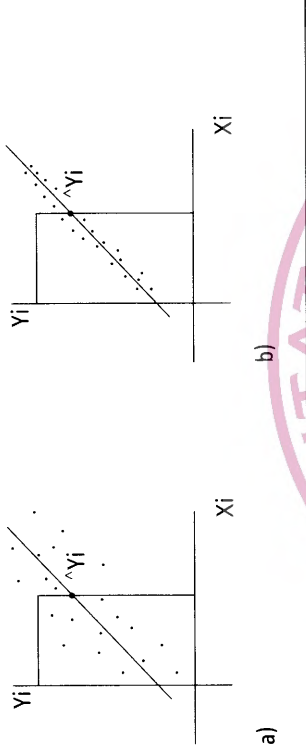
A chiusura di questa elementare trattazione della regressione lineare, è necessario porre l'attenzione sulla necessità di accompagnare la lettura dei risultati forniti dall'applicazione del modello (cioè delle stime dei singoli parametri) con una valutazione di quanto il modello si adatti ai dati empirici. Ci limitiamo a considerare uno dei più utilizzati criteri di valutazione della bontà di adattamento della retta di regressione ai dati, costituito dal coefficiente di determinazione R^2 .

In certi casi il modello rappresenta in modo semplificato i dati empirici senza discostarsi troppo da essi (un caso per la verità piuttosto raro nella ricerca sociale); in altri, il modello approssima la realtà empirica solo a grandi linee. Nei due diagrammi di dispersione di fig. 4, la retta interpolante è identica (stessa intercetta e stessa inclinazione); è però evidente che la capacità di rappresentare la realtà è maggiore nel secondo caso: i punti sono per così dire "maggiormente attratti" dalla retta interpolante.

Abbiamo già visto nella parte di statistica descrittiva (cfr. par. 4.6 del cap. 3) che per valutare la concordanza tra due variabili cardinali si utilizza l'indice r di Pearson. Abbiamo anche detto che il quadrato di tale indice, denominato coefficiente di determinazione, fornisce un valore interpretabile come varianza spiegata di una variabile da parte dell'altra.

...co di R^2 per la
valutazione della
bontà di
adattamento del
modello ai dati

FIGURA 4 Gradi diversi di adattamento della retta interpolante



Ora, alla luce di quanto detto sulla regressione, possiamo dare giustificazione a tale affermazione a partire dal caso della regressione lineare semplice.

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{\text{dev}(\hat{Y})}{\text{dev}(Y)} = \frac{\text{var}(\hat{Y})}{\text{var}(Y)} = \frac{\text{var}(a + bX)}{\text{var}(Y)} = \frac{\text{var}(a) + \text{var}(bX)}{\text{var}(Y)} = \frac{\text{var}(a) + b^2 \text{var}(X)}{\text{var}(Y)}$$

Quest'ultima espressione si può indicare con β_x^2 , ricordando che

$$\beta_x = b_x \cdot \frac{s_x}{s_y}$$

Ma poiché β_x è uguale nella regressione semplice all'indice di correlazione r_{xy} , ne consegue che $R^2 = r_{xy}^2$.

L'equivalenza non è più valida qualora si passi alla regressione multipla. R^2 può essere utilizzato anche nella regressione multipla, nel qual caso assume però il significato di **coefficiente di determinazione multiplo**: esso rappresenta la capacità complessiva di un gruppo di variabili indipendenti di spiegare la varianza della variabile dipendente (per ulteriori approfondimenti sulla correlazione multipla e la sua interpretazione si rinvia a Blacklock, 1960, trad. it. pp. 600 ss.). In generale, R^2 è interpretabile, nella regressione, come il quadrato della correlazione tra la variabile Y (ovvero la variabile dipendente osservata) e la variabile $\hat{Y}$ (cioè la variabile dipendente riprodotta dal modello).

R^2 e r^2 sono equivalenti solo nella regressione semplice

Introduzione alla statistica per la ricerca sociale

Dunque, più in generale, R^2 è una misura della capacità della retta di regressione di rappresentare in modo parsimonioso la nube di punti visualizzata nel diagramma di dispersione. Quanto più i punti tendono ad allontanarsi dalla retta, tanto più il coefficiente di determinazione tende a zero; quanto più tendono a essere “attratti” dalla retta (e al limite ad allinearsi sopra di essa), tanto più il coefficiente di determinazione tende al valore di 1.

È necessario ancora una volta ricordare che R^2 va usato in modo opportuno: un coefficiente di determinazione vicino a zero indica l'assenza di una relazione di forma lineare tra le variabili, non di una qualsiasi forma di relazione.

Inoltre, se stiamo analizzando un campione probabilistico a fini inferenziali, essa non è una misura appropriata in quanto tende a sovrastimare la bontà di adattamento del modello alla popolazione.

In tal caso è preferibile ricorrere all' R^2 corretto, ovvero il corrispettivo nella regressione di η^2 (che abbiamo incontrato nell'ANOVA): la formula di R^2 corretto tiene conto della dimensione del campione e del numero di stimatori in modo tale da correggere appunto la distorsione (ottimistica) dell' R^2 non corretto.

R^2 corretto è
l'equivalente di η^2
nella regressione

In questo capitolo

1. Con le procedure di inferenza su due variabili si entra in un campo della statistica di notevole interesse pratico della ricerca sociale; in particolare esse risultano utili quando si voglia comparare l'andamento di un fenomeno in due popolazioni distinte. Lo studio dell'inferenza bivariata è inoltre propedeutico per approdare alle tecniche di analisi multivariata.
2. L'Analisi della Varianza (ANOVA) è una tecnica sorta in ambito sperimentale per studiare l'impatto di uno o più fattori su una variabile dipendente continua; successivamente, il suo utilizzo è stato esteso anche all'indagine non sperimentale. Il suo utilizzo più frequente è come test di ipotesi della differenza tra le medie di k campioni indipendenti; essa può comunque anche essere utilizzata come metodo di stima di parametri. La sua applicazione rigorosa è condizionata al verificarsi di una serie di condizioni (assunti); essa si mostra comunque robusta rispetto alla violazione di alcuni assunti e pertanto viene utilizzata anche in condizioni non proprio ottimali.
3. La regressione lineare è un modello che permette di spiegare l'andamento di una variabile dipendente cardinale attraverso una o più variabili esplicative quando si può assumere che la relazione sia sostanzialmente rettilinea. La sua importanza è dovuta anche al fatto che l'equazione e gli assunti su cui si basa sono spesso gli stessi di altre più sofisticate tecniche di analisi multivariata dei dati.

Esercizi e test per l'autoverifica

1. Si vuole sottoporre a test l'ipotesi che le persone più ansiose siano più introverso di quelle meno ansiose. A due gruppi differenziati per livello di ansia viene somministrata una Scala di introversione, il cui punteggio totale va da 0 (molto estroverso) a 20 (molto introverso); i risultati ottenuti sono i seguenti:

ansia bassa	15	20	12	13	12	16	20	20	18
ansia elevata	13	13	17	12	11	18	14	14	10

Si proceda quindi:

- a verificare se le varianze delle due popolazioni da cui sono stati estratti i campioni sono omogenee ($\alpha = 5\%$);
 - a verificare con apposito test di differenza tra le medie, se la differenza tra i due gruppi è significativa ($\alpha = 1\%$).
2. Si consideri la seguente matrice di correlazioni tra cinque variabili relative a giustificabilità di comportamenti in un campione rappresentativo dei cittadini statunitensi (dati tratti dalla *World Value Survey* del 1995).

	V1	V2	V3	V4
V2	0.48 (n = 1506)			
V3	0.26 (n = 1515)	0.26 (n = 1505)		
V4	0.06 (n = 1416)	0.10 (n = 1408)	0.05 (n = 1418)	
V5	0.05 (n = 1466)	0.08 (n = 1457)	0.06 (n = 1467)	0.54 (n = 1393)

Legenda: V1: ottenere benefici statali senza averne diritto; V2: non pagare il biglietto sui trasporti pubblici; V3: accettare bu-
starelle nel proprio lavoro; V4: avere relazioni omosessuali; V5: abortire.

Calcolare quali sono le correlazioni significativamente diverse da 0 a un livello di α del 5%.

3. Calcolare il coefficiente b di regressione semplice di Y su X, sapendo che si tratta di una relazione positiva e che $VAR(X) = 33,92$; $VAR(Y) = 6,50$; $R^2 = 0,94$.
4. Quale tra le seguenti vv.aa. viene utilizzata per il test della differenza tra le medie di due piccoli campioni indipendenti?
- a) χ^2
 - b) normale non standardizzata
 - c) normale standardizzata
 - d) t di Student

5. Il test di Wilcoxon serve a valutare la significatività della differenza tra:
- a) due medie in campioni dipendenti b) due mediane in campioni indipendenti
 - c) due medie in campioni indipendenti d) due varianze in campioni indipendenti
6. Nella regressione lineare multipla, R^2 è utilizzabile come misura di:
- a) intensità dell'effetto di X su Y b) adattamento del modello ai dati
 - c) intensità dell'effetto di Y su X d) significatività delle stime

