

ANNA PAOLA ERCOLANI
ALESSANDRA ARENI
LUIGI LEONE

Statistica per la psicologia

II. Statistica inferenziale
e analisi dei dati



L'analisi della varianza

L'ultimo capitolo di questo volume è dedicato ad una tecnica di elaborazione dei dati attuabile solo se la pianificazione della ricerca è stata fatta non solo in modo accurato, cercando di controllare tutte le possibili variabili di disturbo, ma anche tenendo specificamente presente questo particolare tipo di elaborazione. Questa tecnica, chiamata *analisi della varianza*, sostanzialmente affronta problemi di confronto tra le medie di due (analogamente alla *t* di Student) o più campioni tenendo conto contemporaneamente di più variabili. Inizieremo ovviamente col trattare i casi più semplici, dopo aver discusso i principi generali del metodo, ma affronteremo anche casi più complessi, come l'analisi della varianza di tipo fattoriale e l'analisi della covarianza; tratteremo inoltre un modello di analisi della varianza non parametrico.

1. PRINCIPI GENERALI

Il termine «*analisi della varianza*», pur trattandosi in realtà di un'analisi di medie, deriva dal fatto che tale tecnica statistica si basa sulla scomposizione della variabilità totale dei dati del problema in due parti: **variabilità sperimentale**, dovuta alle variabili introdotte o studiate dal ricercatore, e **variabilità accidentale** o **residua**, dovuta a tutte quelle variabili che non è possibile controllare in una ricerca (o che non sono state controllate).

Iniziamo con un esempio per facilitare al lettore la comprensione di tali principi: supponiamo di voler verificare se «istruzioni impartite in modo amichevole inducano nei Ss prestazioni migliori».

Per poter verificare tale ipotesi dobbiamo organizzare un piano di ricerca nel quale la variabile *sperimentale* «tipo di istruzioni impartite» abbia almeno due modalità o due livelli: la modalità che interessa al ricercatore e una modalità «di controllo»: ad esempio, nel nostro caso, «istruzioni impartite in modo amichevole» e «istruzioni impartite in modo neutro».

È chiaro che l'interesse del ricercatore è centrato sulla possibile influenza della variabile sperimentale sulla prestazione dei soggetti: cioè il ricercatore è interessato ad attribuire l'*effetto* «migliore prestazione» alla *causa* «modo amichevole di impartire le istruzioni». È necessario perciò che nell'esperimento sia previsto un qualche «controllo» che consenta di attribuire l'effetto alla causa che interessa e non ad altre: per questa ragione le modalità della variabile sperimentale *devono* essere almeno due. Il pregio essenziale della tecnica che stiamo trattando è quello di poter controllare, a priori, le possibili fonti di variazione, per poi suddividere la variabilità dei dati tra le fonti medesime.

In quanto abbiamo detto possiamo identificare alcuni concetti importanti: innanzitutto quello di variabile sperimentale (o **variabile indipendente**) che viene manipolata e/o dosata dallo sperimentatore che è interessato ad osservarne gli effetti sulla **variabile** cosiddetta **dipendente**, cioè nella *risposta* dei Ss sottoposti ad esperimento. Il ricercatore deve ovviamente essere sicuro che non esistano altre **variabili**, cosiddette **di disturbo**, non controllate, che possono produrre effetti sulla variabile dipendente, effetti che, proprio perché non controllati, potrebbero essere attribuiti erroneamente all'influenza della variabile sperimentale. Infine, tutti gli esperimenti, che hanno come oggetto il comportamento umano, devono tener conto anche della possibile interferenza della cosiddetta **variabile interveniente** costituita dalle caratteristiche dei Ss. Al di là dell'impianto metodologico dell'esperimento, tuttavia, la tecnica di analisi dei dati è la stessa e di questa ci occuperemo principalmente: i risultati dell'analisi, però, potranno essere interpretati in modo diverso a seconda delle variabili che sono in gioco. Per esempio, dobbiamo distinguere tra **variabili indipendenti sperimentali** e **variabili indipendenti differenziali**: le prime sono effettivamente mani-

polabili dallo sperimentatore che decide sul livello di presenza dello stimolo o del «trattamento»: diversa difficoltà di un compito, dosi diverse di farmaci, modo diverso di guidare un gruppo, ecc., possono essere considerate, infatti, variabili effettivamente sperimentali, mentre sesso, scolarità, classe sociale, età e anche caratteristiche psicologiche (livello intellettuale, socievolezza, autoritarismo, ecc.) sono variabili composite che lo sperimentatore può considerare a diversi livelli senza però essere in grado di scinderle nei singoli, ipotetici, componenti (per esempio, la variabile età trascina fatalmente con sé altre variabili quali modelli culturali, tipo di esperienze, opportunità fisiche, ecc.). In quest'ultimo caso, i risultati dell'analisi non potranno pertanto essere interpretati in termini di causa-effetto, ma solo in termini descrittivi, anche se la tecnica di analisi è perfettamente identica.

Sta perciò al ricercatore interpretare correttamente la sua analisi tenendo conto della qualità delle variabili in gioco ed è in questo senso che l'analisi della varianza è, a nostro avviso, una tecnica «educativa», poiché costringe i ricercatori inesperti a riflettere sull'esperimento *prima* di raccogliere i dati. Il principio generale dell'analisi della varianza si basa, come abbiamo detto, sulla possibilità di scomporre la variabilità totale in due parti: una attribuibile all'influenza della variabile sperimentale e una residua, che racchiude tutta la variabilità non controllata.

Ritorniamo ora al nostro esempio e analizziamo il «ragionamento logico» su cui si basa l'analisi della varianza.

L'ipotesi del ricercatore è dunque che il gruppo di soggetti cui sono state impartite istruzioni in modo amichevole (che viene detto «**gruppo sperimentale**») abbia una prestazione media migliore di quello cui sono state impartite le stesse istruzioni in modo neutro (che viene detto «**gruppo di controllo**»).

Supponiamo di aver effettuato l'esperimento, dopo aver controllato le possibili variabili di disturbo, su 10 Ss assegnati a caso al gruppo sperimentale (5 Ss) e al gruppo di controllo (5 Ss): supponiamo che la nostra variabile dipendente (cioè la prestazione richiesta ai Ss) sia la quantità di parole di una lista ricordate dopo 5 minuti; mentre la variabile indipendente sperimentale è il diverso modo di dare le istruzioni ai Ss.

Supponiamo ancora di aver ottenuto i risultati esposti nella tabella 4.1.

TAB. 4.1. Dati grezzi e medie del Gruppo Sperimentale (Gr.S.) e del Gruppo di Controllo (Gr.C.)

VARIABILE INDIPENDENTE			
«MODO AMICHEVOLE» (Gr. S.)		«MODO NEUTRO» (Gr. C.)	
Ss	N. parole ricordate	Ss	N. parole ricordate
1	10	6	8
2	10	7	7
3	14	8	10
4	16	9	14
5	15	10	11
$\bar{x}_S$	13	$\bar{x}_C$	10

L'ipotesi statistica che vogliamo verificare è:

$$H_0: \mu_S = \mu_C$$

$$H_1: \mu_S > \mu_C$$

dove con μ_S e μ_C vengono indicate le medie che avrebbero le popolazioni ipotetiche di tutti i Ss trattati in «modo amichevole» (μ_S), e di tutti i Ss trattati in «modo neutro» (μ_C).

TAB. 4.2. Variazione «totale» dei dati della tabella 4.1

DATI GREZZI		$x_{ij} - \bar{x}$		$(x_{ij} - \bar{x})^2$	
Gr. S.	Gr. C.	Gr. S.	Gr. C.	Gr. S.	Gr. C.
10	8	10 - 11.5 = -1.5	8 - 11.5 = -3.5	(-1.5) ² = 2.25	(-3.5) ² = 12.25
10	7	10 - 11.5 = -1.5	7 - 11.5 = -4.5	(-1.5) ² = 2.25	(-4.5) ² = 20.25
14	10	14 - 11.5 = 2.5	10 - 11.5 = -1.5	(2.5) ² = 6.25	(-1.5) ² = 2.25
16	14	16 - 11.5 = 4.5	14 - 11.5 = 2.5	(4.5) ² = 20.25	(2.5) ² = 6.25
15	11	15 - 11.5 = 3.5	11 - 11.5 = .5	(3.5) ² = 12.25	(0.5) ² = .25
65	50			43.25	41.25
$T = 115$		gdl = $n_S + n_C - 1 =$		$\Sigma = 43.25 + 41.25 =$	
$\bar{x}_T = 11.5$		$= 10 - 1$		$= 84.50$	

Nel caso sia vera l'ipotesi nulla (la variabile sperimentale non influenza la prestazione), i due campioni si possono considerare come provenienti dalla stessa popolazione, cioè i due gruppi di 5 Ss si possono considerare come un unico campione del quale calcoleremo la media e la varianza. I calcoli per la variazione totale sono indicati nella tabella 4.2.

Si può osservare che la variazione totale deriva da due diverse fonti di variazione: l'una costituita dalle differenze che esistono tra il gruppo sperimentale e quello di controllo, cioè *tra i gruppi* e l'altra dalle differenze che esistono tra gli individui in ciascun gruppo, cioè *entro i gruppi*. Proviamo ora a calcolare separatamente la variabilità derivata dalle due fonti di variazione individuate.

La **variazione «tra i gruppi»**, detta anche **between** o **tra i trattamenti**, si può isolare eliminando la variabilità esistente all'interno di ciascun gruppo, cioè rendendo tutti gli individui dello stesso gruppo uguali tra loro ed uguali alla media del gruppo. In questo modo non si modifica la variabilità tra i gruppi, ma si annulla la variabilità all'interno dei gruppi: nella tabella 4.2a viene esemplificato il modo di procedere.

Analogamente possiamo calcolare la **variazione «entro i gruppi»**, detta anche **within** o di **errore** o **residua**, isolandola dalla variazione tra i gruppi, cioè tenendo conto *solo* della variazione all'interno di ciascun gruppo. La variazione verrà così calcolata sugli scarti dalla media di ciascun gruppo. Nella tabella 4.2b viene semplificato il modo di procedere.

Sulla base dei risultati ottenuti nelle tabelle 4.2, 4.2a e 4.2b possiamo controllare che la variabilità totale è stata scomposta in due parti corrispondenti alle due fonti di variazione citate: in realtà la scomposizione non riguarda la varianza vera e propria, intesa come somma dei quadrati degli scarti diviso il numero di gdl, ma solo la somma dei quadrati degli scarti (devianza) e, separatamente i gdl:

$$84.50 = 22.50 + 62$$

$$9 = 1 + 8$$

Le varianze vengono calcolate dividendo le devianze per i rispettivi gdl.

Si può dimostrare¹ che, se è vera l'ipotesi nulla e, quindi, i due campioni provengono dalla stessa popolazione, la varianza *fra i gruppi* e la varianza

TAB. 4.2a. Variazione «tra i gruppi» dei dati della tabella 4.1

DATI GREZZI		DATI GREZZI SENZA VARIANZA ENTRO I GRUPPI (x_i)		$x_i - \bar{x}$		$(x_i - \bar{x})^2$	
GR. S.	GR. C.	GR. S.	GR. C.	GR. S.	GR. C.		
10	8	13	10	$13 - 11.5 = 1.5$	$10 - 11.5 = -1.5$	$(1.5)^2 = 2.25$	$(-1.5)^2 = 2.25$
10	7	13	10	$13 - 11.5 = 1.5$	$10 - 11.5 = -1.5$	$(1.5)^2 = 2.25$	$(-1.5)^2 = 2.25$
14	10	13	10	$13 - 11.5 = 1.5$	$10 - 11.5 = -1.5$	$(1.5)^2 = 2.25$	$(-1.5)^2 = 2.25$
16	14	13	10	$13 - 11.5 = 1.5$	$10 - 11.5 = -1.5$	$(1.5)^2 = 2.25$	$(-1.5)^2 = 2.25$
15	11	13	10	$13 - 11.5 = 1.5$	$10 - 11.5 = -1.5$	$(1.5)^2 = 2.25$	$(-1.5)^2 = 2.25$
65	50	65	50			11.25	11.25
$\bar{x}_S = 13$ $\bar{x}_C = 10$ $gdl = k - 1 =$				$\Sigma = 11.25 + 11.25 =$			
$\bar{x}_T = 11.5$ $= 2 - 1 = 1$				$= 22.50$			

TAB. 4.2b. Variazione «entro i gruppi» dei dati della tabella 4.1

DATI GREZZI		$(x_i - \bar{x}_S)$		$(x_i - \bar{x}_C)$		$(x_i - \bar{x}_S)^2$		$(x_i - \bar{x}_C)^2$	
GR. S.	GR. C.	GR. S.	GR. C.	GR. S.	GR. C.				
10	8	$10 - 13 = -3$	$8 - 10 = -2$	$(-3)^2 = 9$	$(-2)^2 = 4$				
10	7	$10 - 13 = -3$	$7 - 10 = -3$	$(-3)^2 = 9$	$(-3)^2 = 9$				
14	10	$14 - 13 = 1$	$10 - 10 = 0$	$(1)^2 = 1$	$0 = 0$				
16	14	$16 - 13 = 3$	$14 - 10 = 4$	$(3)^2 = 9$	$(4)^2 = 16$				
15	11	$15 - 13 = 2$	$11 - 10 = 1$	$(2)^2 = 4$	$(1)^2 = 1$				
65	50			32	30				
$\bar{x}_S = 13$ $\bar{x}_C = 10$		$gdl = (n_S - 1) + (n_C - 1) =$		$\Sigma = 32 + 30 = 62$					
		$= (5 - 1) + (5 - 1) = 8$							

entro i gruppi sono due stime indipendenti della stessa varianza: la varianza della popolazione; tra queste due stime la varianza *tra i gruppi* tende ad essere maggiore della varianza *entro i gruppi*. Con il test statistico F , di cui è nota la distribuzione campionaria, potremo perciò verificare l'ipotesi nulla: se

$$F = \frac{\text{varianza tra i gruppi}}{\text{varianza entro i gruppi}}$$

è superiore al valore di F critico in corrispondenza ad un livello prefissato di probabilità e ai gdl delle due varianze, allora potremo respingere l'ipotesi nulla e accettare l'ipotesi alternativa; infatti, attraverso la verifica dell'uguaglianza tra le varianze, possiamo verificare indirettamente l'uguaglianza tra le medie, poiché se è vera l'ipotesi H_0 , e le due varianze sono stime della varianza della popolazione, i campioni provengono dalla stessa popolazione e le medie sono stime di un'unica media, la media della popolazione.

Il principio di scomposizione della devianza su cui si basa l'analisi della varianza è molto generale: infatti possiamo avere sia più campioni ($k > 2$) sia più fonti di variazione, come vedremo in seguito su modelli più complessi. Ci interessa qui che il lettore abbia chiaro il significato implicito della tecnica: esistono infatti, come vedremo, formule semplificate che consentono di eseguire facilmente il calcolo ma che non sono così esplicitamente legate alla logica del procedimento.

Vanno tuttavia considerate alcune condizioni per l'applicabilità dell'analisi della varianza: la prima e più importante è l'**omoschedasticità**, cioè la condizione che le popolazioni da cui si suppone provengano i campioni abbiano la stessa varianza. La necessità di questa condizione è facilmente intuibile ed è già stata discussa trattando il test t di Student; comunque si può qui aggiungere che una maggiore varianza di una popolazione porterebbe ad una minore precisione della media relativa: in questo caso la variabilità in gioco non sarebbe più omogenea dando peso diverso, per esempio, alla variabilità sperimentale rispetto a quella residua o viceversa.

Un'ulteriore importante condizione riguarda l'**addittività** degli effetti della variabile sperimentale e di altre possibili variabili controllate: si deve cioè poter pensare che la variabile sperimentale *aggiunge* qualcosa alla condizione base; il peso, per esempio, aumenta in modo moltiplicativo e non additivo: così, se l'aumento di peso prodotto dalla variabile sperimentale è del 10%, usando Ss di peso diverso (A di 70 kg e B di 50 kg), otteniamo aumenti di peso diversi: A aumenterà di 7 kg e B di 5 risultando perciò che B è aumentato meno di A. Se la condizione di addittività non è rispettata, in genere la varianza residua o entro i gruppi tende ad essere grande, limitando così la significatività di F .

In genere il ricercatore sa se gli effetti non sono additivi e può trasformare i

suoi dati per riportare tali effetti all'additività: le trasformazioni più classiche sono quella logaritmica e quella in radice quadrata².

2. L'ANALISI DELLA VARIANZA A UNA VIA: PROCEDIMENTO DI CALCOLO

Il procedimento di calcolo per l'analisi della varianza ricorre a formule complesse che trattano matematicamente quanto detto nel paragrafo 1. Per facilitare la comprensione delle formule e delle dimostrazioni che seguono, definiamo dapprima la **matrice dei dati** e la simbologia cui fare riferimento. Definiamo una generica matrice $n \times k$ costituita da un insieme di dati, x_{ij} , disposti su n righe e k colonne, dove k è il numero dei gruppi da confrontare; x_{ij} è quindi la misura della variabile dipendente relativa all' i -esimo soggetto dell' j -esimo gruppo. Ovviamente n può essere anche diverso da un gruppo all'altro.

TAB. 4.3. Rappresentazione simbolica di una matrice di dati

GRUPPI				
1	2	...	j	k
x_{11}	x_{12}	...	x_{1j}	x_{1k}
x_{21}	x_{22}	...		
.	.	...		
.	.	...		
x_{i1}	x_{i2}	...	x_{ij}	x_{ik}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
x_{n1}	x_{n2}	...	x_{nj}	x_{nk}
T_1	T_2		T_j	T_k
Q_1	Q_2		Q_j	Q_k
n_1	n_2		n_j	n_k
$\bar{x}_1$	$\bar{x}_2$		$\bar{x}_j$	$\bar{x}_k$

dove:

n_j = numerosità dell' j -esimo gruppo e $\sum_{j=1}^k n_j = N$

$$T_j = \sum_{i=1}^{n_j} x_{ij} \text{ (totale per colonna)}$$

$$T = \sum_{j=1}^k T_j \text{ (totale generale)}$$

$$Q_j = \sum_{i=1}^{n_j} x_{ij}^2 \text{ (totale degli } x^2 \text{ per colonna)}$$

$$Q = \sum_{j=1}^k Q_j \text{ (totale generale degli } x^2 \text{)}$$

$$\bar{x}_j = \text{media dell}'j\text{-esimo gruppo}$$

$$\bar{x} = \text{media generale}$$

La **devianza totale** (cfr. tab. 4.2) espressa in simboli sarà:

$$\text{devianza totale} = \sum_{i=1}^{n_j} \sum_{j=1}^k (x_{ij} - \bar{x})^2$$

$$\text{nell'esempio: } \sum_{i=1}^5 \sum_{j=1}^2 (x_{ij} - 11.5)^2 = 84.50$$

Tale devianza può essere scomposta in due parti, come abbiamo visto, e cioè (cfr. tabb. 4.2a e 4.2b):

devianza totale = devianza tra gruppi + devianza entro i gruppi o residua

$$\sum_{i=1}^{n_j} \sum_{j=1}^k (x_{ij} - \bar{x})^2 = \sum_{j=1}^k \sum_{i=1}^{n_j} (\bar{x}_j - \bar{x})^2 + \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$$

Per una dimostrazione matematica di questa relazione vedi più sotto. Consideriamo uno scarto qualunque dalla media generale ($x_{ij} - \bar{x}$); aggiungendo e sottraendo la quantità $\bar{x}_j$, il valore dello scarto non cambia:

$$(x_{ij} - \bar{x}) = (x_{ij} - \bar{x}_j + \bar{x}_j - \bar{x}) = (x_{ij} - \bar{x}_j) + (\bar{x}_j - \bar{x})$$

Eleviamo al quadrato i due membri dell'uguaglianza e consideriamo tutti gli scarti al quadrato del gruppo j -esimo (sommando quindi rispetto a i):

$$\sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2 = \sum_{i=1}^{n_j} [(x_{ij} - \bar{x}_j) + (\bar{x}_j - \bar{x})]^2$$

Il secondo membro dell'uguaglianza è assimilabile al quadrato di un binomio in cui $a = (x_{ij} - \bar{x}_j)$ e $b = (\bar{x}_j - \bar{x})$. Perciò ricordando che $(a + b)^2 = a^2 + 2ab + b^2$, potremo scrivere:

$$\begin{aligned}\sum_1^{n_j} (x_{ij} - \bar{x})^2 &= \sum_1^{n_j} [(x_{ij} - \bar{x}_j)^2 + 2(x_{ij} - \bar{x}_j)(\bar{x}_j - \bar{x}) + (\bar{x}_j - \bar{x})^2] = \\ &= \sum_1^{n_j} (x_{ij} - \bar{x}_j)^2 + 2\sum_1^{n_j} (x_{ij} - \bar{x}_j)(\bar{x}_j - \bar{x}) + \sum_1^{n_j} (\bar{x}_j - \bar{x})^2\end{aligned}$$

Trattandosi di sommatoria rispetto a i il termine $(\bar{x}_j - \bar{x})$ è costante e si può portare fuori dalla sommatoria:

$$\sum_1^{n_j} (x_{ij} - \bar{x})^2 = \sum_1^{n_j} (x_{ij} - \bar{x}_j)^2 + 2(\bar{x}_j - \bar{x}) \sum_1^{n_j} (x_{ij} - \bar{x}_j) + \sum_1^{n_j} (\bar{x}_j - \bar{x})^2$$

ma $\sum_1^{n_j} (x_{ij} - \bar{x}_j) = 0$ perché è la somma degli scarti della media per il gruppo j -esimo.

Pertanto $2(\bar{x}_j - \bar{x}) \sum_1^{n_j} (x_{ij} - \bar{x}_j) = 0$ e quindi

$$\sum_1^{n_j} (x_{ij} - \bar{x})^2 = \sum_1^{n_j} (x_{ij} - \bar{x}_j)^2 + \sum_1^{n_j} (\bar{x}_j - \bar{x})^2$$

Sommando ora rispetto a tutti i gruppi si ottiene:

$$\sum_1^k \sum_1^{n_j} (x_{ij} - \bar{x})^2 = \sum_1^k \sum_1^{n_j} (x_{ij} - \bar{x}_j)^2 + \sum_1^k \sum_1^{n_j} (\bar{x}_j - \bar{x})^2$$

$$\text{devianza totale} = \text{devianza residua} + \text{devianza tra i gruppi}$$

I gradi di libertà relativi a ciascuna devianza godono della stessa proprietà additiva; pertanto:

$$\begin{array}{ccccc}\text{gdl della devianza} & = & \text{gdl della devianza} & + & \text{gdl della devianza} \\ \text{totale} & & \text{tra i gruppi} & & \text{residua} \\ N - 1 & = & (k - 1) & + & (N - k)\end{array}$$

Più precisamente: i gdl della devianza *totale* sono dati dal numero totale dei soggetti meno un vincolo, che è la media generale; i gdl della devianza *tra i gruppi* sono dati dal numero di medie dei gruppi (che corrisponde al numero dei gruppi) meno un vincolo, che è la media generale; i gdl della devianza *residua* sono dati dal numero totale di soggetti meno i k vincoli rappresentati dalle k medie dei gruppi.

Nell'esempio citato avremo:

$$\begin{array}{rclclcl}\text{devianza totale} & = & \text{devianza tra i gruppi} & + & \text{devianza residua} \\ 84.50 & = & 22.50 & + & 62 \\ \text{gdl } 9 & = & 1 & + & 8\end{array}$$

Le varianze saranno¹:

$$\text{varianza totale} = \frac{\sum_j \sum_i (x_{ij} - \bar{x})^2}{N - 1} = \frac{84.50}{9}$$

$$\text{varianza tra i gruppi} = \frac{\sum_j \sum_i (\bar{x}_j - \bar{x})^2}{k - 1} = \frac{22.50}{1}$$

$$\text{varianza residua} = \frac{\sum_j \sum_i (x_{ij} - \bar{x}_j)^2}{N - k} = \frac{62}{8}$$

Le formule appena discusse riguardano ovviamente $k \geq 2$ gruppi. L'ipotesi nulla relativa a k gruppi sarà:

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k,$$

mentre l'ipotesi alternativa sarà:

$$H_1: \text{almeno una } \mu_j \text{ maggiore delle altre.}$$

Procediamo alla verifica delle ipotesi costruendo la tabella riassuntiva classica dell'analisi della varianza.

TAB. 4.4. Tabella riassuntiva dell'analisi della varianza per k gruppi e N soggetti

FONTI DI VARIANZA	DEVIANZA	GDL	VARIANZA	CONFRONTI
Tra i gruppi	$\sum_j \sum_i (\bar{x}_j - \bar{x})^2$	$k - 1$	$\frac{\sum_j (\bar{x}_j - \bar{x})^2}{k - 1}$	$F = \frac{\text{var. tra i gr.}}{\text{var. resid.}}$
Residua	$\sum_j \sum_i (x_{ij} - \bar{x}_j)^2$	$N - k$	$\frac{\sum_j \sum_i (x_{ij} - \bar{x}_j)^2}{N - k}$	
Totale	$\sum_j \sum_i (x_{ij} - \bar{x})^2$	$N - 1$		

Per i dati dell'esempio in tabella 4.1, lo schema riassuntivo sarà:

TAB. 4.5. Tabella riassuntiva dell'analisi della varianza per 2 gruppi

FONTI DI VARIANZA	DEVIANZA	GDL	VARIANZA	CONFRONTI
Tra i gruppi	22.50	1	22.50	$F = \frac{22.50}{7.75} = 2.90$
Residua	62	8	7.75	
Totale	84.50	9		

Fissato $\alpha = .05$, sulla tavola D in corrispondenza a $(k - 1)$ e $(N - k)$ gdl (nell'esempio in corrispondenza a 1 e 8 gdl) troviamo $F_{\text{critico}} = 5.32$. Poiché $F < F_{\text{critico}}$, accettiamo l'ipotesi nulla.

Nel quadro 4.1 viene riportato un esempio di analisi della varianza per più gruppi.

QUADRO 4.1.

Esempio di analisi della varianza per più gruppi

Supponiamo di voler verificare l'ipotesi che il livello di istruzione influisca sull'atteggiamento nei confronti della subordinazione della donna. Consideriamo tre gruppi di soggetti così caratterizzati:

- soggetti con licenza elementare (E);

– soggetti con scuola media superiore (M);

– soggetti con laurea (L);

e misuriamo l'atteggiamento suddetto, che è la variabile dipendente della ricerca, con la scala S/d . Le ipotesi da verificare saranno:

$$H_0: \mu_E = \mu_M = \mu_L \quad \text{e} \quad H_1: \mu_L \text{ maggiore delle altre}$$

Da 18 soggetti, di cui 6 con licenza elementare, 6 con scuola media superiore e 6 con laurea, si siano ottenuti i seguenti risultati:

LIVELLO DI ISTRUZIONE					
E		M		L	
x_1	x_1^2	x_2	x_2^2	x_3	x_3^2
6	36	5	25	1	1
5	25	4	16	4	16
7	49	3	9	2	4
6	36	2	4	3	9
4	16	4	16	4	16
6	36	2	4	2	4
$T_1 = 34$		$T_2 = 20$		$T_3 = 16$	
$\bar{x}_1 = 5.7$		$\bar{x}_2 = 3.3$		$\bar{x}_3 = 2.7$	
$n_1 = 6$		$n_2 = 6$		$n_3 = 6$	
$Q_1 = 198$		$Q_2 = 74$		$Q_3 = 50$	
				$T = 70$	
				$\bar{x} = 3.9$	
				$N = 18$	
				$Q = 322$	

$$\begin{aligned} \text{devianza totale} &= \sum_j \sum_i (x_{ij} - \bar{x})^2 = \\ &= (6 - 3.9)^2 + (5 - 3.9)^2 + \dots + \\ &+ (4 - 3.9)^2 + (2 - 3.9)^2 = 50 \end{aligned}$$

$$\text{gdl} = N - 1 = 18 - 1 = 17$$

$$\begin{aligned} \text{devianza tra i gruppi} &= \sum_i \sum_j (\bar{x}_j - \bar{x})^2 = \\ &= 6 \times (5.7 - 3.9)^2 + 6 \times (3.3 - 3.9)^2 + \\ &+ 6 \times (2.7 - 3.9)^2 = 30 \end{aligned}$$

$$\text{gdl} = k - 1 = 3 - 1 = 2$$

$$\begin{aligned}
 \text{devianza residua} &= \sum_i \sum_j (x_{ij} - \bar{x}_j)^2 = \\
 &= (6 - 5.7)^2 + (5 - 5.7)^2 + \dots + \\
 &+ (6 - 5.7)^2 + (5 - 3.3)^2 + \dots + \\
 &+ (2 - 3.3)^2 + (1 - 2.7)^2 + \dots + \\
 &+ (2 - 2.7)^2 = 20
 \end{aligned}$$

$$\text{gdl} = N - k = 18 - 3 = 15$$

Possiamo verificare empiricamente l'uguaglianza:

$$\begin{array}{rclcl}
 \text{devianza totale} & = & \text{devianza tra i gruppi} & + & \text{devianza residua} \\
 50 & = & 30 & + & 20 \\
 \text{gdl } 17 & = & 2 & + & 15
 \end{array}$$

Tabella riassuntiva dell'analisi della varianza

FONTE DI VARIANZA	DEVIANZA	GDL	VARIANZA	CONFRONTI
Tra i gruppi	30	2	15	$F = \frac{15}{1.33} = 11.28$
Residua	20	15	1.33	
Totale	50	17		

Fissato $\alpha = .01$, sulla tavola D troviamo, con 2 e 15 gdl, $F_{\text{critico}} = 6.36$. Poiché $F > F_{\text{critico}}$, respingiamo l'ipotesi nulla e, osservando le medie, concludiamo che l'istruzione influisce sull'atteggiamento studiato nel senso che a istruzione più elevata corrisponde un punteggio medio di S/d significativamente più basso. (Per un'analisi più approfondita della differenza tra coppie di medie vedi oltre i confronti ortogonali.)

Nella pratica si usano formule abbreviate che rendono più semplici e rapidi i calcoli necessari.

$$\text{Definito } C = \frac{(\sum_j \sum_i x_{ij})^2}{N} = \frac{T_j^2}{N} \text{ (detto «termine di correzione»)}$$

si dimostra che:

$$1. \text{ devianza totale} = Q - C$$

$$2. \text{ devianza tra i gruppi} = \sum_j \frac{T_j^2}{n_j} - C$$

$$3. \text{ devianza residua} = Q - \sum_j \frac{T_j^2}{n_j}$$

Infatti:

$$1. \text{ devianza totale} =$$

$$\begin{aligned}
 &= \sum_i \sum_j (x_{ij} - \bar{x})^2 = \sum_i \sum_j (x_{ij}^2 - 2\bar{x}x_{ij} + \bar{x}^2) = \\
 &= \sum_i \sum_j x_{ij}^2 - 2\sum_i \sum_j \bar{x}x_{ij} + \sum_j \sum_i \bar{x}^2
 \end{aligned}$$

ma, essendo $\bar{x}$ un valore costante, potremo scrivere:

$$\sum_i \sum_j (x_{ij} - \bar{x})^2 = \sum \sum x_{ij}^2 - 2\bar{x} \sum \sum x_{ij} + N\bar{x}^2$$

Ricordando che $\bar{x} = \frac{\sum \sum x_{ij}}{N}$, avremo:

$$\begin{aligned}
 &= \sum_i \sum_j x_{ij}^2 - 2 \frac{\sum_i \sum_j x_{ij}}{N} \sum_i \sum_j x_{ij} + N \frac{(\sum_i \sum_j x_{ij})^2}{N} = \\
 &= \sum_i \sum_j x_{ij}^2 - 2 \frac{(\sum_i \sum_j x_{ij})^2}{N} + \frac{(\sum_i \sum_j x_{ij})^2}{N} = \\
 &= \sum_i \sum_j x_{ij}^2 - \frac{(\sum_i \sum_j x_{ij})^2}{N}
 \end{aligned}$$

che, usando la simbologia della tabella 4.3, equivale a:

$$\sum_i \sum_j (x_{ij} - \bar{x})^2 = Q - C$$

$$2. \text{ devianza residua} = \sum_i \sum_j (x_{ij} - \bar{x}_j)^2$$

Considerando la somma degli scarti al quadrato per il solo gruppo j -esimo e sviluppando il quadrato con il procedimento usato al punto 1, avremo:

$$\sum_i (x_{ij} - \bar{x}_j)^2 = \sum_i x_{ij}^2 - \frac{(\sum_i x_{ij})^2}{n_j} = Q_j - \frac{T_j^2}{n_j}$$

Considerando ora tutti i k gruppi (cioè sommando rispetto a j):

$$\begin{aligned}\sum_j \sum_i (x_{ij} - \bar{x}_j)^2 &= \left(Q_1 - \frac{T_1^2}{n_1} \right) + \left(Q_2 - \frac{T_2^2}{n_2} \right) + \dots + \left(Q_k - \frac{T_k^2}{n_k} \right) = \\ &= Q_1 + Q_2 + \dots + Q_k - \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \dots + \frac{T_k^2}{n_k} \right) = \sum_j Q_j - \sum_j \frac{T_j^2}{n_j} = Q - \sum_j \frac{T_j^2}{n_j}\end{aligned}$$

3. devianza *tra i gruppi* = devianza *totale* – devianza *residua*

$$Q - C - \left(Q - \sum_j \frac{T_j^2}{n_j} \right) = \frac{\sum_j T_j^2}{n_j} - C$$

QUADRO 4.2.

*Esempi di analisi della varianza:
uso delle formule abbreviate*

Esempio 1. Riprendendo l'esempio del quadro 4.1 calcoliamo le devianze utilizzando le formule abbreviate.
Per comodità riportiamo i calcoli intermedi:

LIVELLO DI ISTRUZIONE			TOTALI
E	M	L	
$T_1 = 34$	$T_2 = 20$	$T_3 = 16$	$T = 70$
$Q_1 = 198$	$Q_2 = 74$	$Q_3 = 50$	$Q = 322$
$n_1 = 6$	$n_2 = 6$	$n_3 = 6$	$N = 18$

I calcoli necessari sono:

$$\text{il «termine di correzione» } C = \frac{T^2}{N} = \frac{70^2}{18} = 272$$

• devianza *totale* =

$$= Q - C = 322 - 272 = 50 \quad \text{gdl} = 17$$

• devianza *tra i gruppi* =

$$\begin{aligned}&= \sum_j \frac{T_j^2}{n_j} - C = \left(\frac{34^2}{6} + \frac{20^2}{6} + \frac{16^2}{6} \right) - 272 = \\ &= 302 - 272 = 30 \quad \text{gdl} = 2\end{aligned}$$

• devianza *residua* = devianza *totale* – devianza *tra i gruppi* =

$$= 50 - 30 = 20$$

oppure:

• devianza *residua* =

$$\begin{aligned}&= Q - \sum_j \frac{T_j^2}{n_j} = 322 - \left(\frac{34^2}{6} + \frac{20^2}{6} + \frac{16^2}{6} \right) = \\ &= 322 - 302 = 20 \quad \text{gdl} = 15\end{aligned}$$

I risultati sono identici a quelli ottenuti nel quadro 4.1 con il procedimento di calcolo esteso.

Esempio 2. Supponiamo di voler verificare l'ipotesi che il *tipo di lavoro* influisca sul livello di *stress* del lavoratore. Considerando 3 tipi di lavoratori con diverso livello di intervento personale nel processo di decisione e precisamente G3 = *manager* (alto livello); G2 = *quadri* (livello medio); G1 = *stagisti* (basso livello) e definendo x = misura dello *stress* rilevata mediante un apposito strumento, le ipotesi da verificare saranno:

$$H_0: \mu_1 = \mu_2 = \mu_3$$

$$H_1: \mu_1 < \mu_2 < \mu_3$$

Consideriamo 18 soggetti di cui 6 *stagisti*, 5 *quadri* e 7 *manager* e misuriamo la variabile x = *stress* (a punteggio elevato corrisponde alto livello di *stress*).

G1		G2		G3	
x_1	x_1^2	x_2	x_2^2	x_3	x_3^2
40	1600	57	3249	53	2809
41	1681	40	1600	50	2500
26	676	51	2601	56	3136
36	1296	60	3600	56	3136
21	441	65	4225	54	2916
23	529			75	5625
				70	4900
$T_1 = 187$		$T_2 = 274$	$T_3 = 424$	$T = 885$	
$Q_1 = 6223$		$Q_2 = 15275$	$Q_3 = 26122$	$Q = 47620$	
$\bar{x}_1 = 31.17$		$\bar{x}_2 = 54.80$	$\bar{x}_3 = 60.57$	$\bar{x} = 49.17$	
$n_1 = 6$		$n_2 = 5$	$n_3 = 7$	$N = 18$	

I calcoli necessari sono:

$$C = \frac{885^2}{18} = 43512.5$$

- devianza totale:

$$= 47620 - 43512.5 = 4107.5$$

$$\text{gdl} = 18 - 1 = 17$$

- devianza tra i gruppi =

$$= \frac{187^2}{6} + \frac{274^2}{5} + \frac{424^2}{7} - 43512.5 = 3013.15$$

$$\text{gdl} = 3 - 1 = 2$$

- devianza residua =

$$= 4107.5 - 3013.15 = 1094.35$$

$$\text{gdl} = 17 - 2 = 15$$

Lo schema riassuntivo dell'analisi della varianza sarà:

FONTI DI VARIABILITÀ	DEVIANZA	GDL	VARIANZA	CONFRONTI
Tra i gruppi	3013.15	2	1506.57	$F = \frac{1506.57}{72.96} = 20.65$
Residua	1094.35	15	72.96	
Totale	4107.50	17		

Fissato $\alpha = .01$, per 2 e 15 gdl sulla tavola D troviamo $F_{\text{critico}} = 6.36$. Poiché $F > F_{\text{critico}}$, respingiamo l'ipotesi nulla e concludiamo che il tipo di lavoro influenza il livello di stress (per un'analisi più approfondita vedi quadro 4.4 sui confronti ortogonali).

2.1. Confronto tra analisi della varianza e t di Student

Nel capitolo 3 per la verifica delle ipotesi sulla differenza tra le medie nel caso di due campioni indipendenti, si era utilizzato il test t di Student. Alla luce di quanto introdotto in questo capitolo, possiamo tuttavia considerare il caso di due campioni come caso particolare di analisi della varianza. Per $k = 2$ le due tecniche (analisi della varianza e t di Student) sono equivalenti: si dimostra, infatti, che $t = \sqrt{F}$. È tuttavia consigliabile usare l'analisi della varianza, perché tale tecnica costringe lo sperimentatore a organizzare in modo più preciso il piano degli esperimenti. L'esempio seguente dimostrerà l'assoluta equivalenza tra analisi della varianza nel caso $k = 2$ e t di Student.

QUADRO 4.3.

*Equivalenza tra analisi della varianza e t di Student
per la verifica delle ipotesi sulla differenza
tra le medie nel caso di due campioni indipendenti*

Supponiamo di voler verificare l'ipotesi che l'aggiunta di un commento al voto ottenuto in una prova scritta di italiano favorisca un miglioramento nella prova successiva.