

PARTE B

ANALISI DELLA VARIANZA (ANOVA).

L'obiettivo generale dell'analisi della varianza (ANOVA) è quello di verificare la presenza di differenze significative tra le medie di differenti popolazioni.

Per realizzare tale obiettivo si analizzano campioni diversi di osservazioni e si verifica l'effetto indotto dalla definizione di gruppi differenti sulla variabilità complessiva

N.B

$$\text{VARIANZA POPOLAZIONE: } \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n} = SS/n$$

SS=DEVIANZA=Sum of Squares

Facile Esempio: Scomposizione della somma dei quadrati (SS)

	Gruppo1	Gruppo2
Osservazione 1	2	6
Osservazione 2	3	7
Osservazione 3	1	5
Totale	6	18
Media	2	6
Somma dei Quadrati (SS)	2	2
Media Complessiva	4	Totale complessivo 24
Somma dei Quadrati Totale	28	

L'SS TOTALE (28) SI PUÒ PARTIZIONARE

nell'SS dovuta alla variabilità *entro i gruppi*

SS(E) → **DEVIANZA WITHIN**

es . $2+2=4$

e

nell'SS dovuta alla variabilità *tra i gruppi*

SS(B) → **DEVIANZA BETWEEN**

es . $3((2-4)^2+(6-4)^2)=24$

3=numero di osservazioni per gruppo.....

$$SS(Totale)=SS(E)+SS(B)=4+24=28$$

In generale

La *variabilità entro gruppi* non si può facilmente individuare nel disegno corrente e quindi viene calcolata per differenza tra la variabilità totale e la variabilità tra i gruppi. Per questo motivo è chiamata **variabilità dell'errore**

La *variabilità tra i gruppi* è l'effetto della distinzione in gruppi. Per questo motivo è chiamata **variabilità dell'effetto**

Dalla scomposizione della devianza posso sapere quanta parte della variabilità totale è dovuta all'effetto e quanta parte è dovuta all'errore ma solo per i campioni esaminati

Analisi della varianza

Test di significatività sulla differenza tra le medie delle popolazioni da cui sono estratti i campioni

$H_0: \tau_1 = \tau_2$ (medie delle popolazioni da cui sono stati estratti i $k=2$ gruppi)

TEST

$$F = \frac{SS(B)/(k-1)}{SS(E)/(n-k)} = \frac{MS(B)}{MS(E)} = \frac{24/(2-1)}{4/(6-2)} = \frac{24}{1} = 24$$

dove $k-1$ sono i gradi di libertà relativi alla stima della variabilità tra i gruppi e $n-k$ i gradi di libertà relativi alla stima dell'errore

Sotto l'ipotesi nulla (che non vi sia differenza tra le medie tra i gruppi della popolazione) ci si aspetterebbe che la varianza stimata sulla base della variabilità entro gruppi dovrebbe essere all'incirca pari a quella dovuta alla variabilità tra gruppi.

Si possono confrontare queste due stime di varianza tramite il test F, che verifica se il rapporto delle due stime delle varianze risulta significativamente maggiore di 1.

Nell'esempio precedente, questo test risulta altamente significativo, infatti F teorico per $\alpha=0,05$ con $(k-1)$ e $(n-k)$ gradi di libertà è uguale a 7,71 quindi si conclude che le medie per i due gruppi sono significativamente differenti l'una dall'altra.

Riassunto in formule: Analisi della varianza a un fattore

consideriamo t_j campioni o trattamenti differenti $j=1, \dots, k$ e n_j osservazioni Y_{ij} per ciascun gruppo

si indichi con $\bar{Y}_{.j}$ e con $T_{.j}$ rispettivamente la media e il totale del j -esimo gruppo

si indichi con $\bar{Y}_{..}$ e con $T_{..}$ rispettivamente la media e il totale generale

Variabilità	gradi di libertà	SS	MS
tra i trattamenti	$k-1$	$\sum_{j=1}^k n_j (\bar{Y}_{.j} - \bar{Y}_{..})^2 = \sum_{j=1}^k \frac{T_{.j}^2}{n_j} - \frac{T_{..}^2}{n}$	$SS(B)/k-1$
errore	$n-k$	$\sum_{j=1}^k \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{.j})^2 = \sum_{j=1}^k \sum_{i=1}^{n_j} Y_{ij}^2 - \sum_{j=1}^k \frac{T_{.j}^2}{n_j}$	$SS(E)/n-k$
Totale	$n-1$	$\sum_{j=1}^k \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{..})^2 = \sum_{j=1}^k \sum_{i=1}^{n_j} Y_{ij}^2 - \sum_{j=1}^k \frac{T_{.j}^2}{n}$	

$$F = \frac{\text{Test}}{SS(E)/(n-k)} = \frac{SS(B)/(k-1)}{MS(E)} = f_{\text{oss}}$$

$$F_{k-1, n-k, \alpha} = f_{\text{teorico}}$$

Se $f_{\text{oss}} > f_{\text{teorico}}$ il test è significativo, le medie sono significativamente differenti

Test sulle medie- Metodo dei contrasti ortogonali

Scelta a priori del numero di confronti da esaminare nel caso in cui $k > 2$

E' piuttosto importante se si verifica una significativa differenza tra le medie dei diversi trattamenti capire quali sono le medie che differiscono tra di loro. Se la scelta dei confronti avviene prima di effettuare l'esperimento diventa importante effettuarla con cura tenendo anche conto del fatto che il numero dei confronti non può eccedere il numero dei gradi di libertà ($k-1$).

Alcune definizioni e notazioni

Si definisce contrasto la combinazione lineare:

$$C_m = c_{1m}T_{.1} + c_{2m}T_{.2} + \dots + c_{jm}T_{.j} + \dots + c_{km}T_{.k}$$

dove :

$$c_{1m} + c_{2m} + \dots + c_{jm} + \dots + c_{km} = 0$$

In presenza di più contrasti è preferibile che i confronti siano indipendenti l'uno dall'altro così test di ipotesi indipendenti possono essere eseguiti

Due contrasti C_m e C_q sono detti ortogonali se:

$$\sum_{j=1}^k c_{jm}c_{jq} = 0$$

La SS di un contrasto è:
$$SS_{C_m} = \frac{C_m^2}{n \sum_{j=1}^k c_{jm}^2}$$

N.B Se il numero di osservazioni per gruppo non è uguale (n) è necessario considerare nelle formule n_j

Esempio : Analisi della varianza e metodo dei contrasti

Esempio 8: Effetto sulla conduttività di 4 differenti tipi di rivestimento di un tubo catodico

I	II	III	IV
56	64	45	42
55	61	46	39
62	50	45	45
59	55	39	43
60	56	43	41

Competamente randomizzato, n=20

Si procede ad un'analisi della varianza sottraendo precedentemente a ciascun valore 50

	I	II	III	IV
	6	14	-5	-8
	5	11	-4	-11
	12	0	-5	-5
	9	5	-11	-7
	10	6	-7	-9
T_j	42	36	-32	-40
n_j	5	5	5	5
$\sum_{j=1}^k \sum_{i=1}^{n_j} Y_{ij}^2 = 1340 \quad \mathbf{T_{..}=6}$ $\mathbf{N=20}$				

$$SS(TOTALE) = \sum_{j=1}^k \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{..})^2 = \sum_{j=1}^k \sum_{i=1}^{n_j} Y_{ij}^2 - \sum_{j=1}^k \frac{T_j^2}{n} = 1340 - \frac{6^2}{20} = 1338,2$$

$$SS(B) = \sum_{j=1}^k n_j (\bar{Y}_{.j} - \bar{Y}_{..})^2 = \sum_{j=1}^k \frac{T_{.j}^2}{n_j} - \frac{T_{..}^2}{n} =$$

$$\frac{42^2}{5} + \frac{36^2}{5} + \frac{-32^2}{5} + \frac{-40^2}{5} + \frac{6^2}{20} = 1135,0$$

$$SS(E) = SS(Totale) - SS(B) = 1338,2 - 1135,0 = 203,2$$

Variabilità	gradi di libertà	SS	MS
tra i trattamenti	3	1135,0	378,3
errore	16	203,2	12,7
Totale	19	1338,2	

$$F_{oss} = 29,8 \quad F_{teorico} = F_{3,16,0,01} = 5,29$$

test significativo

Un analisi dei contrasti può essere effettuata considerando al massimo 3 contrasti (K-1 gradi di libertà)

$$C_1 = T_1 - T_4$$

$$C_2 = T_2 - T_3$$

$$C_3 = T_1 - T_2 - T_3 + T_4$$

	T ₁	T ₂	T ₃	T ₄
C ₁	+1	0	0	-1
C ₂	0	+1	-1	0
C ₃	+1	-1	-1	+1

$$C_1 = +1(42) + 0(36) + 0(-32) - 1(-40) = 82$$

$$C_2 = +0(42) + 1(36) - 1(-32) + 0(-40) = 68$$

$$C_3 = +1(42) - 1(36) - 1(-32) + 1(-40) = -2$$

$$SS_{C_1} = \frac{82^2}{5(2)} = \frac{6724}{10} = 672,4$$

$$SS_{C_2} = \frac{68^2}{5(2)} = \frac{4624}{10} = 462,4$$

$$SS_{C_3} = \frac{-2^2}{5(4)} = \frac{4}{20} = 0,2$$

$$H_1:t_1=t_4 \quad F = \frac{672,4/1}{12,7} = 52,9 \quad \text{significativo } F_{1,16,0,05}=4,49$$

$$H_2:t_2=t_3 \quad F = \frac{464,4/1}{12,7} = 36,4 \quad \text{significativo } F_{1,16,0,05}=4,49$$

$$H_0:t_1+t_4 = t_2+t_3 \quad F = \frac{0,2/1}{12,7} = 0,016 \quad \text{non significativo } F_{1,16,0,05}=4,49$$

Assunto

Nell'analisi della varianza si assume che le varianze entro le popolazioni differenti del disegno siano identiche; questo assunto è detto di omogeneità delle varianze.

Si ricordi che all'inizio si è calcolata la varianza dell'errore sommando tra loro i quadrati degli scarti entro ciascun gruppo. Se le varianze nei due gruppi sono differenti le une dalle altre, allora sommarle assieme non risulta appropriato, e non si ottiene una stima della varianza comune entro i gruppi

Effetti nel violare gli assunti.

Lindman (1974, p. 33) dimostra che la statistica F risulta estremamente robusta rispetto alla violazione di questo assunto

Caso speciale: medie e varianze correlate: si immagini di avere 8 celle nel disegno, 7 con medie all'incirca uguali e una con la media molto più elevata. La statistica F può suggerire in questo caso un effetto statisticamente significativo. Tuttavia, si supponga che vi sia anche una varianza molto più elevata nella cella con la media più elevata, cioè, che le medie e le varianze siano correlate nelle celle. In tal caso, la media maggiore nella singola cella potrebbe in realtà non essere differente, a causa della varianza elevata. Comunque, poiché la statistica F complessiva si basa su una stima aggregata della varianza entro celle, la media più elevata potrà essere identificata come significativamente differente dalle altre, quando nei fatti non risulterebbe tale se si basasse il test sulla varianza entro questa sola cella.

Questo "motivo" -- una media e una varianza elevate in una cella -- si verifica frequentemente quando vi sono outlier presenti nei dati. Uno o due casi estremi in una cella con solo 10 casi possono distorcere profondamente la media, ed accrescere inoltre in maniera rilevante anche la varianza.