Analisi della varianza

- Prof. Giuseppe Verlato
- Sezione di Epidemiologia e Statistica Medica, Università di Verona

ANALISI DELLA VARIANZA - 1

Abbiamo k gruppi, con un numero variabile di unità statistiche.

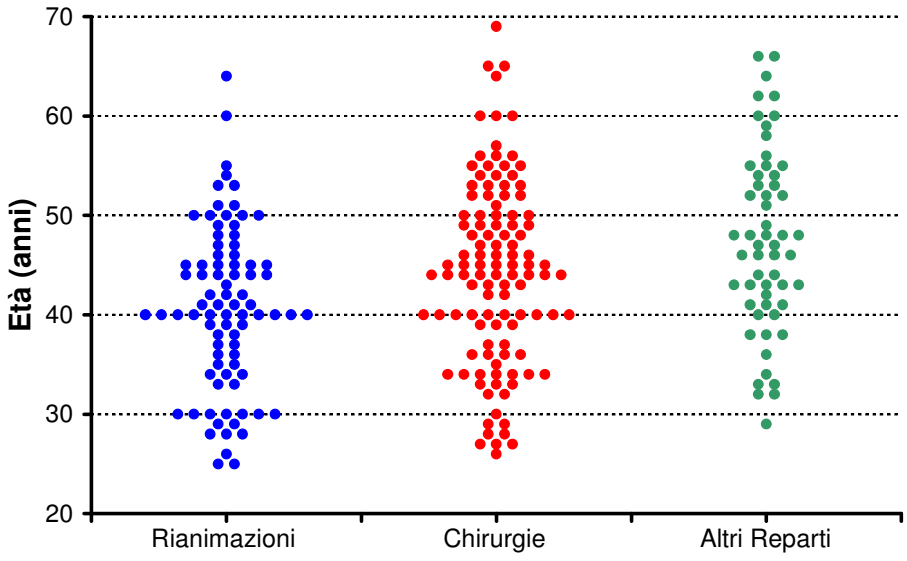
Nella notazione classica, ogni unità statistica viene individuata da due numeri in posizione pedice: il primo indica il gruppo di appartenenza, e il secondo indica la posizione del soggetto all'interno del gruppo.

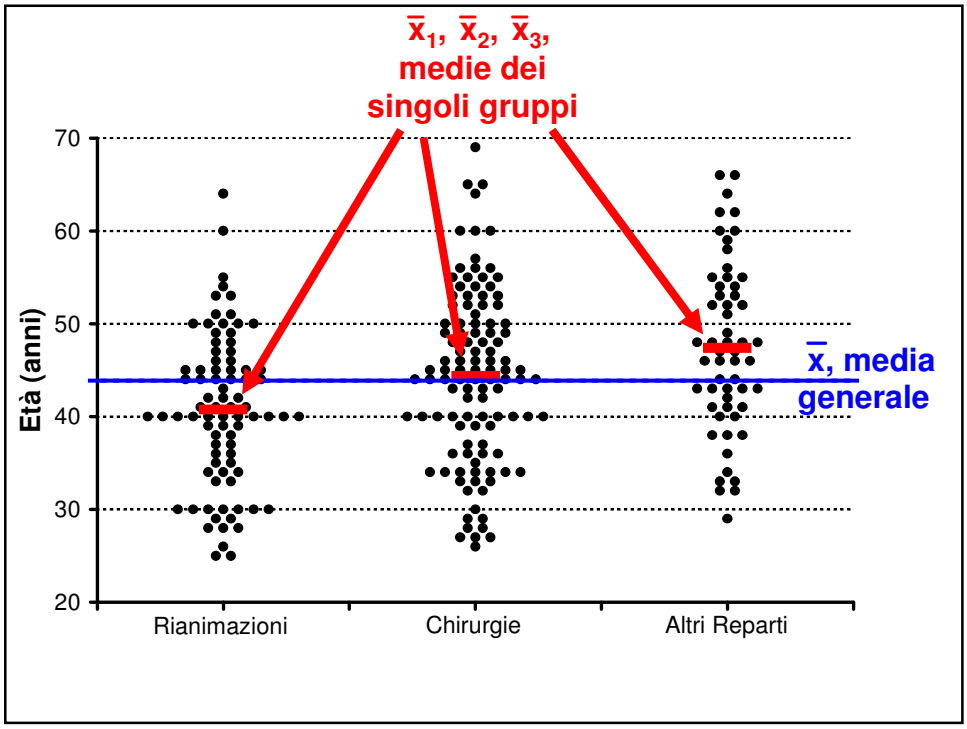
gruppo 1	**gruppo 2**	**gruppo 3**	**……..**	**gruppo k**
x_{11}	x_{21}	x_{31}	……..	x_{k1}
x_{12}	x_{22}	x_{32}	……..	x_{k2}
x_{13}	x_{23}	x_{33}	……..	x_{k3}
x_{14}	x_{24}	x_{34}	……..	x_{k4}
x_{15}	x_{25}	x_{35}	……..	x_{k5}
x_{16}	x_{26}	x_{36}	……..	x_{k6}
x_{17}	x_{27}	x_{37}	……..	x_{k7}
x_{18}		x_{38}	……..	x_{k8}
		x_{39}	……..	x_{k9}
$\overline{x}_{1.}$	$\overline{x}_{2.}$	$\overline{x}_{3.}$	……..	$\overline{x}_{k.}$

ANALISI DELLA VARIANZA - 2

Oltre ad una media generale, $\bar{x}$, abbiamo k medie, una per ognuno dei singoli gruppi, $\bar{x}_{1.}$, $\bar{x}_{2.}$, $\bar{x}_{3.}$, …., $\bar{x}_{k.}$

Età dei medici operanti nelle Rianimazioni, Chirurgie ed altri Reparti di un'Azienda Ospedaliera del Veneto





ANALISI DELLA VARIANZA - 3

Ipotesi $\begin{cases} H_0: \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_0 \\ H_1: \text{almeno una media differisce dalle altre} \end{cases}$

Per rispondere a questa domanda, possiamo fare tante t di Student, confrontando tutte le possibili coppie di medie?

NO, perché altrimenti avremmo un'inflazione (aumento abnorme) di α (alfa), probabilità di errore del I tipo.

Se eseguo 20 t di Student, ciascuna con un α nominale di 0,05, per effetto del caso almeno un test mi risulta significativo:

$\alpha_{effettivo} = 1 - (1 - \alpha_{nominale})^n$, dove n = numero di test effettuati

Ad esempio con 6 t-test e con un $\alpha_{nominale}$ di 0,05,

$\alpha_{effettivo} = 1 - (1 - 0{,}05)^6 = 1 - 0{,}95^6 = 1 - 0{,}735 = 0{,}265$

ANALISI DELLA VARIANZA - 4

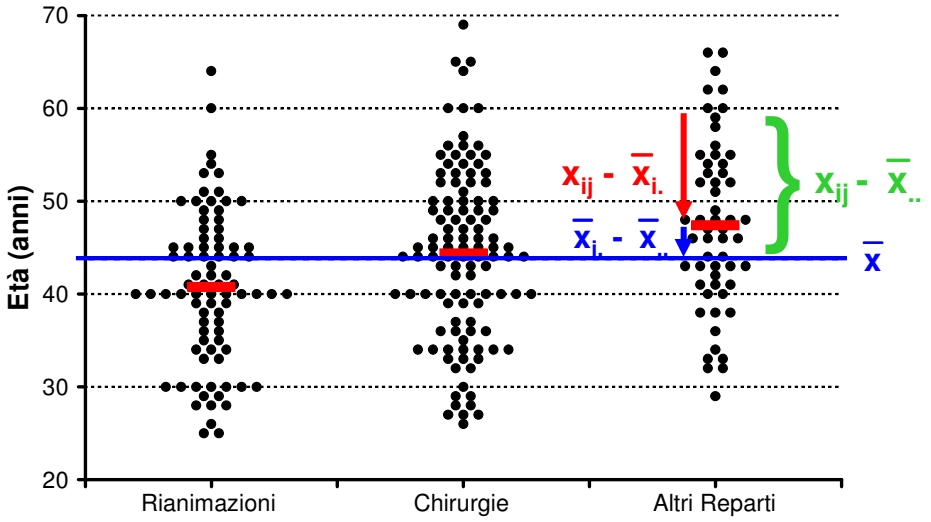
$\alpha_{corretto} = 1 - (1- \alpha_{nominale})^{1/n} = 1 - \sqrt[n]{(1- \alpha_{nominale})}$

Ad esempio con 6 t-test e con un $\alpha_{nominale}$ di 0,05,

$\alpha_{corretto} = 1 - (1- 0,05)^{1/6} = 1 - 0,95^{1/6} = 1 - 0,9915 = 0,0085$

E' meglio quindi ricorrere ad un test globale, che confronti fra di loro tutti i gruppi: **l'analisi della varianza**.

SCOMPOSIZIONE DELLA DEVIANZA nell'Analisi della Varianza - 1



$x_{ij} - \bar{x}_{..}$ = scarto di una singola osservazione (valore j*esimo* del gruppo i*esimo*) dalla media generale

$\bar{x}_{i.} - \bar{x}_{..}$ = scarto della media del gruppo i*esimo* dalla media generale

$x_{ij} - \bar{x}_{i.}$ = scarto di una singola osservazione (valore j*esimo* del gruppo i*esimo*) dalla media del gruppo i*esimo*

SCOMPOSIZIONE DELLA DEVIANZA nell'Analisi della Varianza - 2

Per una singola osservazione:

Variabilità totale — Variabilità entro gruppi

$$(x_{ij} - \bar{x}_{..}) = (\bar{x}_{i.} - \bar{x}_{..}) + (x_{ij} - \bar{x}_{i.})$$

Variabilità fra gruppi

Si può dimostrare che, per tutte le osservazioni:

Devianza totale, SST — Devianza entro gruppi

$$\sum_{i,j}(x_{ij} - \bar{x}_{..})^2 = \sum_{i,j}(\bar{x}_{i.} - \bar{x}_{..})^2 + \sum_{i,j}(x_{ij} - \bar{x}_{i.})^2$$

Devianza tra gruppi

Scomposizione della devianza: schema dei calcoli

Siano **k** i gruppi da confrontare, con numerosità $n_1, n_2, \ldots, n_i, \ldots n_k$, anche diverse tra loro.

Indichiamo con $T_i = \sum_{j=1}^{n_j} x_{ij}$, **con** $G = \sum_{i=1}^{k} T_i$, **con** $N = \sum_{i=1}^{k} n_i$

gruppo 1	gruppo 2	gruppo 3	…	gruppo k
x_{11}	x_{21}	x_{31}	…	x_{k1}
x_{12}	x_{22}	x_{32}	…	x_{k2}
x_{13}	x_{23}	x_{33}	…	x_{k3}
x_{14}	x_{24}	x_{34}	…	x_{k4}
x_{15}	x_{25}	x_{35}	…	x_{k5}
x_{16}	x_{26}	x_{36}	…	x_{k6}
x_{17}	x_{27}	x_{37}	…	x_{k7}
x_{18}		x_{38}	…	x_{k8}
		x_{39}	…	x_{k9}
Totale	T_1	T_2	T_3	… T_k	**G**

SCOMPOSIZIONE DELLA DEVIANZA nell'Analisi della Varianza - 3

Devianza Totale $= \sum_{i=1}^{k} \sum_{j=1}^{n_j} (x_{ij} - \bar{x}_{..})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_j} x_{ij}^2 - (\sum_{i=1}^{k} \sum_{j=1}^{n_j} x_{ij})^2 / N$

Devianza ENTRO gruppi $= \sum_{i=1}^{k} \sum_{j=1}^{n_j} (x_{ij} - \bar{x}_{i.})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_j} x_{ij}^2 - \sum_{i=1}^{k} (\sum_{j=1}^{n_j} x_{ij})^2 / n_i$

Devianza FRA gruppi $= \sum_{i=1}^{k} \sum_{j=1}^{n_j} (\bar{x}_{i.} - \bar{x}_{..})^2 = \sum_{i=1}^{k} (\sum_{j=1}^{n_j} x_{ij})^2 / n_i - (\sum_{i=1}^{k} \sum_{j=1}^{n_j} x_{ij})^2 / N$

Dalla statistica descrittiva sappiamo che la devianza è pari a:

$$\text{Devianza} = \sum_i (x_i - \bar{x}_{..})^2 = \sum_i x_i^2 - (\sum_i x_i)^2 / N$$

Per calcolare la devianza totale per l'ANOVA, si procede nello stesso modo, semplicemente i dati vanno sommati sia lungo le colonne che lungo le righe.

$$\text{Devianza Totale} = \sum_i \sum_j (x_{ij} - \bar{x}_{..})^2 = \sum_i \sum_j x_{ij}^2 - (\sum_i \sum_j x_{ij})^2 / N$$

La devianza ENTRO gruppi (devianza residua) è pari alla somma algebrica delle devianze dei singoli gruppi.

$$\text{Devianza gruppo 1} = \sum_j (x_{1j} - \bar{x}_{1.})^2 = \sum_j x_{1j}^2 - (\sum_j x_{1j})^2 / n_1$$

$$\text{Devianza gruppo 2} = \sum_j (x_{2j} - \bar{x}_{2.})^2 = \sum_j x_{2j}^2 - (\sum_j x_{2j})^2 / n_2$$

$$\text{Devianza gruppo 3} = \sum_j (x_{3j} - \bar{x}_{3.})^2 = \sum_j x_{3j}^2 - (\sum_j x_{3j})^2 / n_3$$

$$\text{SOMMA devianze} = \sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2 = \sum_i \sum_j x_{ij}^2 - \sum_i (\sum_j x_{ij})^2 / n_i$$

$$\text{Devianza FRA gruppi} = \sum_i \sum_j (\bar{x}_{i.} - \bar{x}_{..})^2 = \sum_i n_i (\bar{x}_{i.} - \bar{x}_{..})^2$$

La devianza fra gruppi si può calcolare come differenza fra la devianza TOTALE e la devianza ENTRO gruppi.

devianza totale - devianza entro =

$$\sum_i \sum_j x_{ij}^2 - (\sum_i \sum_j x_{ij})^2 / N - [\sum_i \sum_j x_{ij}^2 - \sum_i (\sum_j x_{ij})^2 / n_i] =$$

$$\sum_i \sum_j x_{ij}^2 - (\sum_i \sum_j x_{ij})^2 / N - \sum_i \sum_j x_{ij}^2 + \sum_i (\sum_j x_{ij})^2 / n_i =$$

$$\sum_i (\sum_j x_{ij})^2 / n_i - (\sum_i \sum_j x_{ij})^2 / N$$

Inferenza sulle varianze - 1

Assunto n.1: Omoscedasticità della varianza
La varianza è omogenea (uguale) in tutti i k gruppi.

Se è vero questo assunto, è possibile ottenere una stima migliore di tale varianza combinando insieme le stime derivanti da ogni gruppo. La **varianza entro gruppi** o **varianza residua** può essere stimata dividendo la somma delle devianze dei singoli gruppi per la somma dei gradi di libertà dei singoli gruppi.

$$\text{Varianza residua} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2 / (N-k)$$

E' possibile dimostrare che il valore atteso della varianza residua è la varianza comune σ^2.

Inferenza sulle varianze - 2

Dividendo la devianza fra gruppi per un numero di gradi di libertà pari al numero di gruppi meno uno, si ottiene una stima della **varianza fra gruppi**:

$$\text{Varianza fra gruppi} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{x}_{i.} - \bar{x}_{..})^2 / (k-1)$$

Assunto n. 2: Le osservazioni (a rigore gli errori)
sono fra loro indipendenti.

Se è vero questo assunto, il valore atteso della varianza fra gruppi è pari a:

$$E(\text{var fra}) = \sigma^2 + \sum_{i=1}^{k} n_i (\mu_{i.} - \bar{\mu})^2 / (k-1)$$

Se è vera H_0, il valore atteso della varianza fra gruppi vale esattamente σ^2 (la varianza residua); in caso contrario il suo valore è maggiore di σ^2 .

Inferenza sulle varianze – 3

Assunto n.3: in ciascun gruppo le osservazioni sono distribuite in modo gaussiano.

Se è vero questo assunto, è possibile eseguire un test di significatività. Infatti, sotto H_0 la **varianza residua (entro gruppi)** e la **varianza fra gruppi** sono stime indipendenti di σ^2.

Il test più appropriato per confrontare queste due varianze è il test F:

F = **var FRA** / **var ENTRO**

Tale statistica, sotto H_0, segue la distribuzione F di Fisher-Snedecor con (k-1) gradi di libertà al numeratore e (N-k) gradi di libertà al denominatore.

Il valore ottenuto va confrontato con una soglia critica $F_{\alpha,(k-1),(N-k)}$ desunta dalle apposite tavole.

Se $F > F_{\alpha,(k-1),(N-k)}$, si rifiuta H_0 ($P<\alpha$); in caso contrario si accetta H_0.

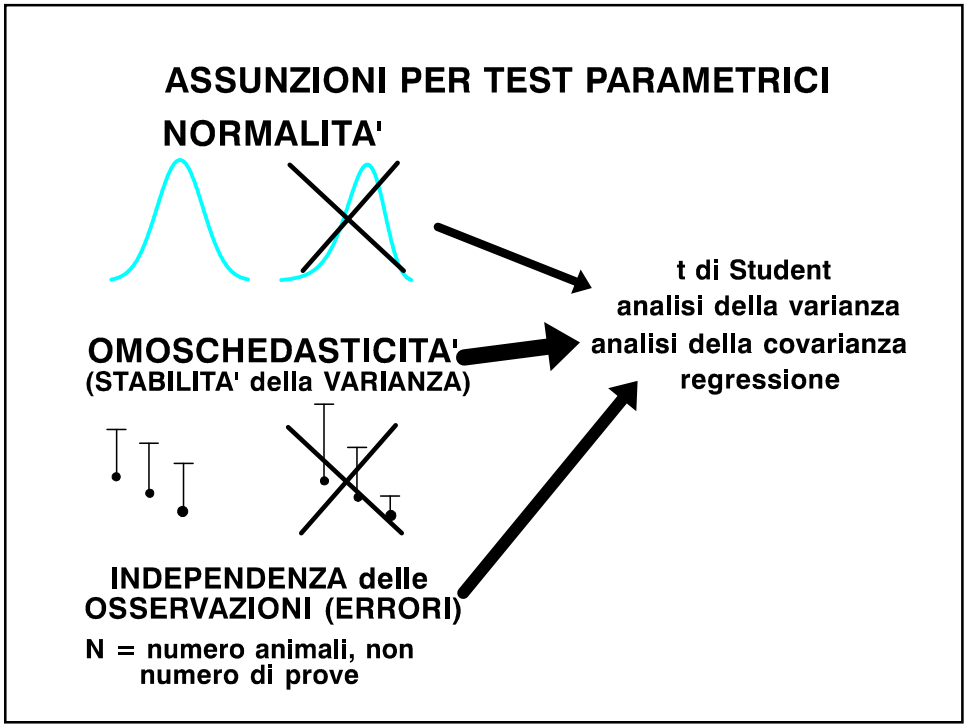


Tabella dell'ANALISI DELLA VARIANZA
a 1 criterio di classificazione

Fonte di variabilita'	Gradi di liberta'	Devianza	Varianza	Test-F
TRA gruppi	k-1	$\sum n_i (\bar{x}_{i.}-\bar{x}_{..})^2$	$DEV_{tra}/(k-1)$	$\frac{VAR_{tra}}{VAR_{entro}}$
ENTRO gruppi (errore)	N-k	$\sum\sum (x_{ij}-\bar{x}_{i.})^2$	$DEV_{entro}/(N-k)$	
Totale	N-1	$\sum\sum (x_{ij}-\bar{x}_{..})^2$		

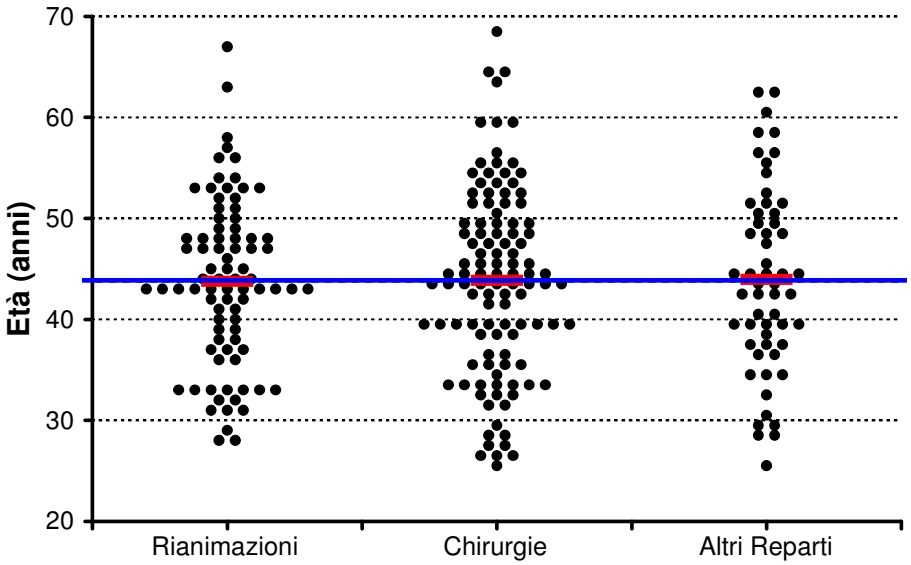
Se H_1 e' vera (almeno 1 media e' diversa dalle altre), allora

$$\sigma^2_{tra} > \sigma^2_{entro} \approx \sigma^2$$

L'età dei medici (n=251) è significativamente diversa nelle Rianimazioni, nelle Chirurgie e negli altri Reparti?

fonte di variabilità	gradi di libertà	devianza	varianza	test F (significatività)
TRA gruppi	2	1546,10	773,05	9,779
ENTRO gruppi	248	19604,73	79,05	(P<0,001)
TOTALE	250	21150,84		

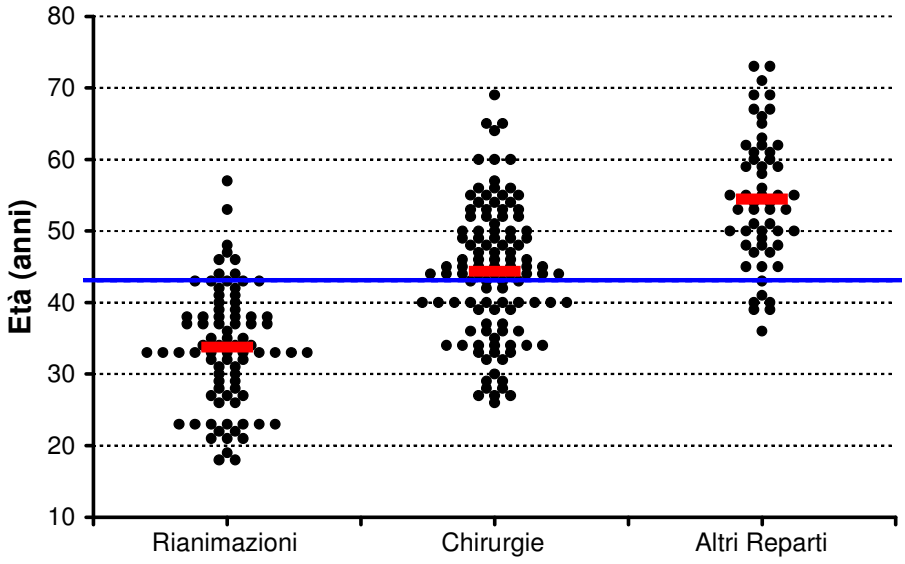
Si rifiuta l'ipotesi nulla: l'età differisce significativamente tra i vari Reparti.



L'età dei medici (n=251) è significativamente diversa nelle Rianimazioni, nelle Chirurgie e negli altri Reparti?

fonte di variabilità	gradi di libertà	devianza	varianza	test F (signi-ficatività)
TRA gruppi	2	1,21	0,60	0,008
ENTRO gruppi	248	19604,73	79,05	(P=0,992)
TOTALE	250	19605,94		

Si accetta l'ipotesi nulla: l'età non differisce significativamente tra i vari Reparti.



L'età dei medici (n=251) è significativamente diversa nelle Rianimazioni, nelle Chirurgie e negli altri Reparti?

fonte di variabilità	gradi di libertà	devianza	varianza	test F (signi-ficatività)
TRA gruppi	2	14628,6	7314,3	92,53
ENTRO gruppi	248	19604,7	79,05	(P<0,001)
TOTALE	250	34233,3		

Si rifiuta l'ipotesi nulla: l'età differisce significativamente tra i vari Reparti.

Se il test F (il test globale) risulta significativo, si possono confrontare fra loro le singole medie, utilizzando dei test appropriati che vanno sotto il nome di confronti o contrasti multipli.

Questo tipo di analisi viene definita anche “post hoc”.

1) I contrasti multipli in genere utilizzano come stima della variabilità casuale la varianza residua, calcolata nell’ambito dell’analisi della varianza.

2) I contrasti multipli sono costruiti in modo tale da evitare l’inflazione di α, probabilità di errore del I tipo.

CONTRASTI (o CONFRONTI) MULTIPLI (POST HOC ANALYSIS)

TEST di SCHEFFE' e' il test piu' conservativo, si confrontano sia medie che gruppi di medie

TEST di TUKEY si confrontano tutte le possibili coppie di medie

TEST di DUNNETT si confrontano tutte le singole medie con un controllo

CORREZIONE di BONFERRONI

Si moltiplica il valore di 'p' (probabilita' di errore del I tipo) per il numero di test effettuati

Ad esempio, se effettuo 3 test statistici:

p iniziale		p corretta
0.03	*3=	0.09
0.15	*3=	0.45
0.01	*3=	0.03