

Analisi della varianza

Il modello di Analisi della varianza è legato al particolare disegno sperimentale scelto

Noi faremo riferimento al *Disegno Completamente Randomizzato* in cui k trattamenti vengono assegnati casualmente ad n unità sperimentali

T_1	T_2	T_3	$\dots$	T_i	$\dots$	T_k
y_{11}	y_{21}	y_{31}	$\dots$	y_{i1}	$\dots$	y_{k1}
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
y_{1n_1}	y_{2n_2}	y_{3n_3}	$\dots$	y_{in_i}	$\dots$	y_{kn_k}
$\bar{y}_{1\cdot}$	$\bar{y}_{2\cdot}$	$\bar{y}_{3\cdot}$	$\dots$	$\bar{y}_{i\cdot}$	$\dots$	$\bar{y}_{k\cdot}$

In questo caso il modello statistico assume la forma

$$\boxed{Y_{ij} = \mu_i + \varepsilon_{ij}} \quad \text{dove } \varepsilon_{ij} \sim N(0, \sigma^2)$$

dove Y_{ij} è il valore della variabile risposta in corrispondenza del trattamento T_i per la j -esima osservazione del gruppo.

Y_{ij} è uguale al valore atteso della variabile risposta μ_i per il trattamento T_i a cui si somma un errore accidentale. La natura casuale dell'errore viene tradotta in probabilità assumendo per ε_{ij} una distribuzione Normale con media nulla.

La varianza dell'errore, σ^2 , misura la variabilità dovuta a fattori casuali comune a tutte le osservazioni

Dalla normalità di ε_{ij} segue la normalità di Y_{ij} ; solo quest'ultima è una variabile aleatoria osservabile poiché non saremo mai in grado di osservare separatamente l'errore casuale contenuto nei nostri dati. Importanti assunzioni sono alla base del modello

- LE UNITÀ ALL'INTERNO DEI GRUPPI SONO OMOGENEE
- EVENTUALI FATTORI DI CONFONDIMENTO SONO STATI CONTROLLATI DALLO SPERIMENTATORE

NE DERIVA CHE, A PARITÀ DI TRATTAMENTO, LA VARIABILITÀ OSSERVATA NELLA RISPOSTA È DI NATURA PURAMENTE CASUALE.

ESPRIMIAMO IL VALORE ATTESO DELLA RISPOSTA
COME

$$E[Y_{ij}] = \mu_i = \eta + \alpha_i \quad \text{per} \quad i = 1, \dots, k$$

DOVE

- $\eta \longrightarrow$ EFFETTO MEDIO GENERALE
- $\alpha_i \longrightarrow$ EFFETTO DELL'IESIMO TRATTAMENTO,
DIFFERENZIALE RISPETTO A η

SIAMO PASSATI DA $k(+1)$ PARAMETRI
A $k+1(+1)$ PARAMETRI

Il modello è sovrapparametrizzato

In realtà esistono tra i parametri dei vincoli *naturali*

$$\begin{aligned}\eta &= \frac{\sum \mu_i}{k} \\ \sum \mu_i &= k\eta \\ \mu_i &= \eta + \alpha_i \\ \sum \mu_i &= k\eta + \sum \alpha_i\end{aligned}$$

$$\begin{aligned}&\longrightarrow \sum \alpha_i = 0 \\ \alpha_k &= -\sum_{i=1}^{k-1} \alpha_i\end{aligned}$$

STIMIAMO L'EFFETTO DEI DIVERSI TRATTAMENTI

Gli stimatori ottimali (sia nel senso dei minimi quadrati sia ottimizzando sulla scala della verosimiglianza) per i parametri originari μ_i sono

$$\hat{\mu}_i = \frac{\sum_j Y_{ij}}{n_i} = \bar{Y}_{i.} \quad i = 1, \dots, k$$

Da cui

$$\begin{cases} \hat{\eta} = \frac{\sum_{ij} Y_{ij}}{n} = \bar{Y}_{..} \text{ dove } n = \sum n_i \\ \hat{\alpha}_i = \bar{Y}_{i.} - \bar{Y}_{..} \quad i = 1, \dots, k \end{cases}$$

SOLUZIONE SEMPLICE ED INTUITIVA

Il modello di riparametrizzazione utilizzato non è l'unico possibile

SE ESISTE UN GRUPPO DI RIFERIMENTO (T_k) POSSIAMO SCEGLIERE UNA DIVERSA CHIAVE DI LETTURA

$$\mu_i = \beta_{(k)} + \beta_i \quad i = 1, \dots, k-1$$

I corrispondenti stimatori saranno

$$\begin{cases} \hat{\mu}_k &= \hat{\beta}_{(k)} = \bar{Y}_k. \\ \hat{\beta}_i &= \hat{\mu}_i - \hat{\beta}_{(k)} = \bar{Y}_{i\cdot} - \bar{Y}_k. \end{cases}$$

Tuttavia la stima dei parametri originari (e delle loro differenze) resta univoca

- $\hat{\mu}_m - \hat{\mu}_l = \hat{\eta} + \hat{\alpha}_m - \hat{\eta} - \hat{\alpha}_l = \bar{Y}_{m\cdot} - \bar{Y}_l.$
- $\hat{\mu}_m - \hat{\mu}_l = \hat{\beta}_{(k)} + \hat{\beta}_m - \hat{\beta}_{(k)} - \hat{\beta}_l = \bar{Y}_{m\cdot} - \bar{Y}_l.$

VERIFICHIAMO IL SISTEMA D'IPOTESI

$$\begin{cases} H_0 & : \mu_1 = \dots = \mu_k = 0 \\ H_1 & : \mu_l \neq \mu_m \text{ per almeno un coppia } (l, m) \end{cases}$$

Rispetto al t-test:

- Estendiamo il confronto da due popolazioni a k popolazioni
- L'ipotesi alternativa è molto ampia
- La costruzione della statistica test si sposta dal confronto tra le medie campionarie all'analisi della variabilità osservata

SCOMPOSIZIONE DELLA DEVIANZA

DA

$$(Y_{ij} - \bar{Y}_{..}) = (Y_{ij} - \bar{Y}_{i.}) + (\bar{Y}_{i.} - \bar{Y}_{..})$$

SEGUE (dopo qualche calcolo ...)

$$\sum_i \sum_j (Y_{ij} - \bar{Y}_{..})^2 = \sum_i \sum_j (Y_{ij} - \bar{Y}_{i.})^2 + \sum_i n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2$$

$$\begin{array}{ccccc} \sum_i \sum_j (Y_{ij} - \bar{Y}_{..})^2 & = & \sum_i \sum_j (Y_{ij} - \bar{Y}_{i.})^2 & + & \sum_i n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2 \\ \text{DEVIANZA} & & \text{DEVIANZA} & & \text{DEVIANZA SPIEGATA} \\ \text{TOTALE} & & \text{RESIDUA} & & \text{DAL MODELLO} \\ (TSS) & & (RSS) & & (MSS) \\ Y_{ij} & = & \varepsilon_{ij} & + & \mu_i \end{array}$$

Scomposizione dei gradi di libertà

$$\boxed{\begin{array}{ccccc} \text{TSS} & = & \text{RSS} & + & \text{MSS} \\ (n-1) & & (n-k) & & (k-1) \end{array}}$$

N.B.

$$\begin{aligned} & 2 \sum_i \sum_j (\bar{Y}_{i.} - \bar{Y}_{..})(Y_{ij} - \bar{Y}_{i.}) \\ = & 2 \sum_i (\bar{Y}_{i.} - \bar{Y}_{..}) \underbrace{\sum_j (Y_{ij} - \bar{Y}_{i.})}_{=0} = 0 \end{aligned}$$

$$\frac{\text{RSS}}{n-k} = \frac{\sum_{ij} (Y_{ij} - \bar{Y}_{i.})^2}{n-k} = \frac{\sum_{ij} (Y_{ij} - \hat{\mu}_i)^2}{n-k} = \hat{\sigma}^2$$

È uno stimatore di σ^2 cioè della variabilità di natura accidentale solo se le osservazioni all'interno dei gruppi sono effettivamente omogenee

FONTE DI VARIAZIONE	SS	G.d.l	MS
MODELLO	$\underbrace{\sum_i n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2}_{MSS}$	$k - 1$	$\frac{\sum_i n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2}{k-1}$
RESIDUA	$\underbrace{\sum_{ij} (Y_{ij} - \bar{Y}_{i.})^2}_{RSS}$	$n - k$	$\frac{\sum_{ij} (Y_{ij} - \bar{Y}_{i.})^2}{n-k}$
TOTALE	$\underbrace{\sum_{ij} (Y_{ij} - \bar{Y}_{..})^2}_{TSS}$	$n - 1$	

La nostra statistica test sarà semplicemente il rapporto tra la varianza spiegata dal modello e quella residua

$$F = \frac{\text{MSS}/(k-1)}{\text{RSS}/(n-k)} \sim_{H_0} \mathbb{F}_{(k-1, n-k)}$$

E la regione di rifiuto del test

$$R = \{F : F \geq F_{(k-1, n-k), (1-\alpha)}\}$$

CONFRONTIAMO TRATTAMENTI DIVERSI

SUPPONIAMO DI AVER RIFIUTATO L'IPOTESI NULLA

→ ESISTE UNA DIFFERENZA SIGNIFICATIVA TRA I
TRATTAMENTI SPERIMENTATI

- QUALI DIFFERISCONO?
- QUALE É IL TRATTAMENTO MIGLIORE?

CONFRONTIAMO T_k CON T_l

$$\begin{cases} H_0 & : \mu_k = \mu_l & (\alpha_k = \alpha_l) \\ H_1 & : \mu_k \neq \mu_l & (\alpha_k \neq \alpha_l) \end{cases}$$

STIMIAMO $(\mu_k - \mu_l)$ CON $(\bar{Y}_{k\cdot} - \bar{Y}_{l\cdot})$

PIÙ IN GENERALE DEFINIAMO COME CONFRONTO LA
QUANTITÀ $\sum c_i \mu_i$ CON $\sum c_i = 0$

STIMIAMO $\sum c_i \mu_i$ CON $\sum c_i \bar{Y}_i$.

DALL'IPOTESI $Y_{ij} \sim N(\mu_i, \sigma^2)$ AVREMO

- $\bar{Y}_{i.} \sim N(\mu_i, \frac{\sigma^2}{n_i})$

- $C = \sum c_i \bar{Y}_{i.} \sim N(\sum c_i \mu_i, \sigma^2 \sum_i \frac{c_i^2}{n_i})$

- SOTTO L'IPOTESI $H_0 : \sum_i c_i \mu_i = 0$

$$t = \frac{\sum c_i \bar{Y}_{i.}}{\hat{\sigma} \sqrt{\sum_i \frac{c_i^2}{n_i}}} \sim t_{n-k}$$

DOVE $\hat{\sigma}^2 = \frac{RSS}{n-k}$

- LA STATISTICA t È EQUIVALENTE AD UNA STATISTICA F , basta elevarla al quadrato

$$F = t^2 = \frac{(\sum_i c_i \bar{Y}_i)^2 / \sum_i \frac{c_i^2}{n_i}}{\frac{\text{RSS}}{n-k}} = \frac{\frac{\text{CSS}}{1}}{\frac{\text{RSS}}{n-k}} \sim \mathbb{F}_{(1, n-k)}$$

CSS è adesso la devianza spiegata dal particolare confronto che stiamo analizzando

- SE CONFRONTIAMO 2 TRATTAMENTI OTTENIAMO IL "CLASSICO" t -TEST

$$T = \frac{\bar{Y}_{k\cdot} - \bar{Y}_{l\cdot}}{\hat{\sigma} \sqrt{\frac{1}{n_k} + \frac{1}{n_l}}} \sim t_{(n-k)}$$

TUTTAVIA ADESSO σ^2 VIENE STIMATO UTILIZZANDO L'INFORMAZIONE CONTENUTA NEI k GRUPPI

(gdl = $\sum n_i - k \neq n_k + n_l - 2$)

VOGLIAMO VALUTARE L'EFFETTO DI DUE SOSTANZE ATTIVE (A e B) SULLA CAPACITÀ D'ATTENZIONE IN INDIVIDUI SANI. CIASCUN SUGGETTO VIENE SOTTOPOSTO AD UN TEST REGISTRANDO IL NUMERO DI ERRORI COMMESSI

	A_1 <i>CONTR.</i>	A_2 <i>A</i>	A_3 <i>B</i>	A_4 $A + B$	Totale
	1	12	12	13	
	8	6	4	14	
	9	10	11	14	
	9	13	7	17	
	7	13	8	11	
	7	13	10	14	
	4	6	12	13	
	9	10	5	14	
$\bar{x}$	6.750	10.375	8.625	13.70	9.875
$\hat{\sigma}^2$	8.214	8.839	9.696	2.786	7.384

TABELLA ANOVA

Fonte di Variaz.	SS	G.d.l.	MS
Modello	212.75	3	70.92
Residua	206.75	28	7.384
Totale	419.5	31	

$$F = \frac{70.92}{7.384} = 9.60 \quad \text{p-value} < 0.001$$

RIFIUTIAMO L'IPOTESI NULLA $\rightarrow$ ESISTE ALMENO UNA DIFFERENZA

3 QUESITI:

- (i) IN MEDIA LE SOSTANZE IN SPERIMENTAZIONE HANNO UN QUALCHE EFFETTO SULLA CAPACITÀ DI ATTENZIONE?
- (ii) IL NUMERO DI ERRORI AUMENTA SE LE SOSTANZE VENGONO SOMMINISTRATE CONTEMPORANEAMENTE?
- (iii) L'EFFETTO DELLE DUE SOSTANZE É SIMILE?

3 CONFRONTI

$$H_0(1) : \mu_1 = (\mu_2 + \mu_3 + \mu_4)/3$$

$$\mu_1 - 1/3\mu_2 - 1/3\mu_3 - 1/3\mu_4 = 0$$

$$H_0(2) : \mu_4 - 1/2\mu_2 - 1/2\mu_3 = 0$$

$$H_0(3) : \mu_2 - \mu_3 = 0$$

NEL CASO DEL PRIMO CONFRONTO AVREMO

$$\begin{aligned} c_1 &= \bar{y}_{1.} - 1/3\bar{y}_{2.} - 1/3\bar{y}_{3.} - 1/3\bar{y}_{4.} = \\ &= -4.167 \end{aligned}$$

$$t = \frac{-4.167}{\sqrt{7.384} \sqrt{\frac{1}{8} + \frac{(-1/3)^2}{8} + \frac{(-1/3)^2}{8} + \frac{(-1/3)^2}{8}}} = -3.759 \quad \longrightarrow \quad \text{p-value} = 0.002$$

- POSSIAMO RIASSUMERE I 3 CONFRONTI PIANIFICATI COME SEGUE

	A_1	A_2	A_3	A_3
$\bar{x}_i$	6.75	10.37	8.62	13.75
c_{i1}	1	$-1/3$	$-1/3$	$-1/3$
c_{i2}	0	$-1/2$	$-1/2$	1
c_{i3}	0	1	-1	0

	C	$\sum_i c_i^2$	CSS	F	t	p
$H_0(1)$	-4.167	$4/3$	104.2	14.1	-3.755	0.002
$H_0(2)$	4.250	$3/2$	96.33	13.0	3.606	0.002
$H_0(3)$	1.750	2	12.25	1.66	1.288	0.210

QUANDO DUE CONFRONTI SONO ORTOGONALI?

$$C_k = \sum c_i \bar{Y}_i. \quad C_l = \sum c_i \bar{Y}_i.$$

$$E(C_k C_l) = E(C_k)E(C_l) \text{ INDIPENDENZA TRA V.A. NORMALI}$$

$$\sum_i \frac{c_{ik} c_{il}}{n_i} \overset{\updownarrow}{=} 0 \text{ INDIP. STATISTICA}$$

$$\text{SE } n_i = n_j \longrightarrow$$

$$\sum c_{ik} c_{il} = 0 \text{ INDIPENDENZA LOGICA}$$

→ IN GENERALE OGNI TEST SULLE DIFFERENZE TRA UN DATO INSIEME DI μ_i SARÀ ORTOGONALE AD OGNI TEST CHE COINVOLGE LA LORO SOMMA

$$H_0(1) : \mu_1 - 1/3(\mu_2 + \mu_3 + \mu_4)$$

$$H_0(2) : \mu_4 - 1/2(\mu_2 + \mu_3)$$

$$H_0(3) : \mu_3 - \mu_2$$

- QUANDO É POSSIBILE I CONFRONTI DOVREBBERO ESSERE ORTOGONALI

(IMPARIAMO A FARE DOMANDE DIVERSE)

OGNI CONFRONTO CI DICE QUALCOSA DI NUOVO

$$\left. \begin{array}{l} H_0(4) : \mu_1 = \mu_4 \\ : \mu_4 - \mu_1 = 0 \end{array} \right\} \text{ORTOGONALE A } H_0(1)?$$

$$H_0(1) : \mu_1 = 1/3\mu_2 + 1/3 + \mu_3 + 1/3\mu_4$$

... STRANI RISULTATI

Immaginiamo i dati osservati siano adesso

	A_1	A_2	A_3	A_4
$\bar{x}_i$	5.50	5.7	5.75	9.00
c_{i1}	3	-1	-1	-1
c_{i4}	-1	0	0	1

	C	CSS	F	p
$H_0(1)$	-4.0	10.67	1.21	0.29
$H_0(4)$	3.5	49.00	5.55	0.026

In verità la contraddizione è solo apparente ma l'interpretazione dei risultati non è immediata

- ABBIAMO IN TOTALE $k - 1$ CONFRONTI ORTOGONALI

N.B.

$$\begin{array}{rcccccc}
 CSS(1) & + & CSS(2) & + & CSS(3) & = & MSS \\
 104.7 & + & 96.33 & + & 12.25 & = & 212.75 \\
 \hline
 \text{g.d.l.} & & 1 & & 1 & & 1 & = & 3
 \end{array}$$

La devianza spiegata dai singoli confronti (se ortogonali) ricostruisce esattamente la devianza spiegata dai trattamenti

- POSSIAMO AVERE UNA F GLOBALE NON SIGNIFICATIVA E DELLE F PARTICOLARI SIGNIFICATIVE

Riducendo l'ampiezza dell'ipotesi alternativa aumenta la potenza del test che focalizza l'attenzione in una particolare direzione

SE CONTINUIAMO A FARE CONFRONTI ALLA FINE
ALMENO UNO SIGNIFICATIVO EMERGE

→ ELEVATA PROBABILITÀ DI COMMITTERE UN
ERRORE DI PRIMA SPECIE

ALCUNI PRINCIPI BASILARI

- I CONFRONTI DEVONO ESSERE PIANIFICATI
- IL LORO NUMERO DEVE ESSERE
COERENTE CON I RELATIVI g.d.l.

Se insistiamo nell'esplorazione di molti confronti dobbiamo operare una correzione per molteplicità. Infatti il valore di α fissato come limite di errore nel singolo test non è sufficiente. L'errore globale, cioè la probabilità di commettere un errore di prima specie in almeno uno dei test (familywise error rate) sarà molto più alta. Facendo 10 confronti ciascuno con $\alpha = 0.05$ la probabilità globale di errore sale a 0.40 (vere alcune assunzioni). Il principio generale è quello di ridurre il valore di α nei singoli test. Quanto?

$$r \text{ CONFRONTI} \longrightarrow \alpha_i = \frac{\alpha}{r} \longrightarrow$$

$$Pr\{\text{UNO O PIÙ CONFRONTI SIGNIFICATIVI} | H_0\} < \alpha$$

ma diventiamo molto conservativi. Meglio rivolgersi ad uno statistico ...