	Task Description	Due Date	Status
☐	ANOVA: l'analisi della varianza		
☐			
☐			
☐			
☐			
☐			
☐			
☐			
☐			
☐			

ANOVA: Analisi della varianza
(ANalysis Of VAriance)

generalizzazione del test del confronto tra due medie al caso di g popolazione

studio degli effetti di un fattore sperimentale (variabile qualitative) su una variabile di risposta (variabile quantitative)

MEMO: studio della dipendenza in media

GRUPPI	CARDINALITA'	MEDIE	DEVIANZE	VARIANZE
1	n_1	$\bar{x}_1$	Dev_1	s_1^2
2	n_2	$\bar{x}_2$	Dev_2	s_2^2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	n_i	$\bar{x}_i = \sum_{j=1}^{n_i} \frac{x_{ij}}{n_i}$	$Dev_i = \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$	$s_i^2 = \frac{Dev_i}{n_i - 1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
G	n_G	$\bar{x}_G$	Dev_G	s_G^2
	n			

$\sigma^2 = \sigma^2_{EST} + \sigma^2_{INT}$ proprietà di decomposizione della varianza

$Dev(X) = Dev_{EST} + Dev_{INT}$ in termini di devianza

EST $\rightsquigarrow$ esterna o tra i gruppi

INT $\rightsquigarrow$ interna o all'interno dei gruppi

$$Dev(X) = \sum_{i=1}^{q} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$$

variabilità complessiva del carattere

$$Dev_{EST} = \sum_{i=1}^{q} (\bar{x}_i - \bar{x})^2 n_i$$

misura della differenza tra i gruppi (tra le medie di gruppo e la media complessiva)

$$Dev_{INT} = \sum_{i=1}^{q} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 = \sum_{i=1}^{q} Dev_i = \sum_{i=1}^{q} s_i^2 (n_i - 1)$$

misura della variabilità interna ai gruppi

Le ipotesi dell'ANOVA:

$H_0: \mu_1 = \mu_2 = \dots = \mu_G$ le G popolazioni hanno la stessa media

H_1: almeno una μ_i è diversa dalle altre

NOTA TECNICA: l'ANOVA assume che le G popolazioni siano normali e che abbiano la stessa varianza

(ipotesi di omoscedasticità)

$$\sigma_1^2 = \sigma_2^2 = \dots = \sigma_G^2 = \sigma^2$$

i due stimatori di σ^2 sono ottenuti sfruttando Dev_{INT} e Dev_{EST}

la verifica di H_0 consiste nel confrontare due stimatori della varianza comune, uno valido sia sotto H_0 che sotto H_1 e l'altro valido solo sotto H_0

- STIMATORE 1 PER σ^2 (valido sia sotto H_0 che sotto H_1)

  ↓

  ciascuna delle varianze campionarie s_i^2 è uno stimatore corretto della varianza della popolazione

  ↓

  a partire dai G stimatori disponibili $s_1^2, s_2^2, \ldots, s_G^2$ si ottiene un unico stimatore che ne è la media pesata:

$$\frac{\sum_{i=1}^{G} s_i^2 (n_i - 1)}{n - G} = \frac{\sum_{i=1}^{G} Dev_i}{n - G} = \frac{Dev_{INT}}{n - G}$$

↓

questo stimatore si ottiene senza sapere nulla sulla verità o sulla falsità di H_0

- STIMATORE 2 PER σ^2 (valido sotto H_0)

↓

sotto H_0 si ha che $\mu_i = \mu \quad \forall i = 1, G$

↓

le G medie campionarie $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_G$ si distribuiscono come una normale con media μ e varianza $\frac{\sigma^2}{n_i}$

↓

la varianza campionaria di questi dati:

$$\bar{S}^2 = \frac{\sum_{i=1}^{G} (\bar{X}_i - \bar{X})^2 n_i}{G-1} = \frac{\text{DevEST}}{G-1}$$

è uno stimatore di σ^2

Da cui si ottiene la statistica test F:

stima corretta di σ^2 sotto H_0 (tende a sovrastimare σ^2 se H_0 è falsa)

$$F = \frac{Dev_{TRA} / (g-1)}{Dev_{ENTRO} / (n-g)} \sim F_{g-1,\, n-g}$$

stima corretta di σ^2 sia sotto H_0 che sotto H_1

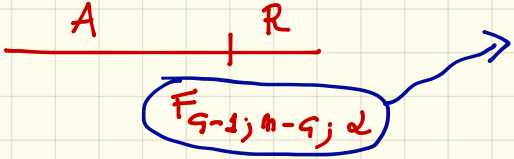
il rapporto tende a essere grande quando H_0 è falsa

(del resto se la Dev_{TRA} è molto alta rispetto alla Dev_{ENTRO} è un segnale che le medie di gruppo sono molto diverse rispetto alla media complessiva)

Le statistiche dell'ANOVA sono di solito riassunte in una tavola di questo tipo:

FONTE DI VARIABILITA'	DEVIANZE (SOMME DEI QUADRATI)	gdl	VARIANZE (MEDIE DEI QUADRATI)	F	p-value
tra i gruppi (spiegata)	Dev_{EST}	$G-1$	$Dev_{EST}/(G-1)$	$\frac{Dev_{EST}/(G-1)}{Dev_{INT}/(n-G)}$	p-value associato alla statistica F
entro i gruppi (residua)	Dev_{INT}	$n-G$	$Dev_{INT}/(n-G)$		
totale	Dev_{TOT}	$n-1$	$Dev_{TOT}/(n-1)$		

La regola di decisione dell'ANOVA è sempre di tipo unilaterale dx:



→ si può procedere confrontando il valore della statistica F con la soglia o confrontando il p-value con l'α

ESERCIZIO (ANOVA)

In una casa editrice ci sono tre redazioni (A, B e C). A partire dalle 3 redazioni sono stati selezionati casualmente 4, 5 e 6 redattori.
L'esperienza in mesi dei redattori è riportata nella seguente tabella:

REDAZIONE	ESPERIENZA
A	25, 27, 29, 31
B	23, 23, 24, 26, 29
C	25, 26, 27, 29, 34, 36

Costruire la tavola ANOVA e, assumendo che le tre popolazioni si distribuiscano normalmente, sottoporre a verifica l'ipotesi nulla che non vi sia differenza tra le esperienze medie nelle tre redazioni al livello di significatività $\alpha = 0.05$.

statistiche di base

REDAZIONE A	REDAZIONE B	REDAZIONE C
$n_A = 4$	$n_B = 5$	$n_C = 6$
$\bar{x}_A = 28$	$\bar{x}_B = 25$	$\bar{x}_C = 29{,}5$
$Dev_A = 20$	$Dev_B = 26$	$Dev_C = 101{,}5$

$$\bar{x} = \frac{28 \times 4 + 25 \times 5 + 29 \times 6}{15} = 27{,}6$$

$$Dev_{TRA} = (28-27{,}6)^2 \times 4 + (25-27{,}6)^2 \times 5 + (29{,}5-27{,}6)^2 \times 6 = 56{,}1$$

$$Dev_{ENTRO} = 20 + 26 + 101{,}5 = 147{,}5$$

$$Dev_{TOT} = Dev_{TRA} + Dev_{ENTRO} = 56{,}1 + 147{,}5 = 203{,}6$$

Fonte di variazione	DEV (SQ)	gdl	VAR (MQ)	F
Tra i gruppi	56,1	2	28.05	2.28
Entro i gruppi	147,5	12	12.29	
TOTALE	203,6	14		

I passi del test ANOVA

① $H_0: \mu_A = \mu_B = \mu_C$

H_1: almeno una μ_i è diversa dalle altre

② $\alpha = 0.05$

③ $F = \frac{MQ_{TRA}}{MQ_{ENTRO}} \sim F_{c-1, n-c}$

④

$F_{2,12;0.05}$

3,89

⑤ $F \in A$

$2,28 < 3,89$

non si rifiuta H_0