

L'Analisi della Varianza

ANOVA

(ANalysis Of VAriance)

ANALISI DELLA VARIANZA

CONCETTI GENERALI

- **Finora abbiamo descritto test di ipotesi finalizzati alla verifica di ipotesi sulla differenza tra parametri di due popolazioni**
- **Spesso si presenta la necessità di prendere in considerazione esperimenti od osservazioni relative a più di due gruppi individuati sulla base di un fattore di interesse**
- **I gruppi sono quindi formati secondo i livelli assunti da un fattore, ad esempio:**
 - **la temperatura di cottura di un oggetto in ceramica che assume diversi livelli numerici come 300°, 350°,400°,450° oppure**
 - **il fornitore che serve una azienda può assumere diversi livelli qualitativi come Fornitore 1, Fornitore 2, Fornitore 3, Fornitore 4**

ANALISI DELLA VARIANZA

- **L'analisi della varianza è una tecnica che consente di confrontare da un punto di vista inferenziale le medie di più di due gruppi (popolazioni)**
- **Quando i gruppi sono definiti sulla base di un singolo fattore si parla di analisi della varianza a un fattore o a una via**
- **Questa procedura, basata su un test F, è una estensione a più gruppi del test t per verificare l'ipotesi sulla differenza tra le medie di due popolazioni indipendenti**
- **Anche se si parla di analisi della varianza in realtà l'oggetto di interesse sono le differenze tra medie nei diversi gruppi e proprio tramite l'analisi della variabilità all'interno dei gruppi e tra gruppi che siamo in grado di trarre delle conclusioni sulla differenza delle medie**

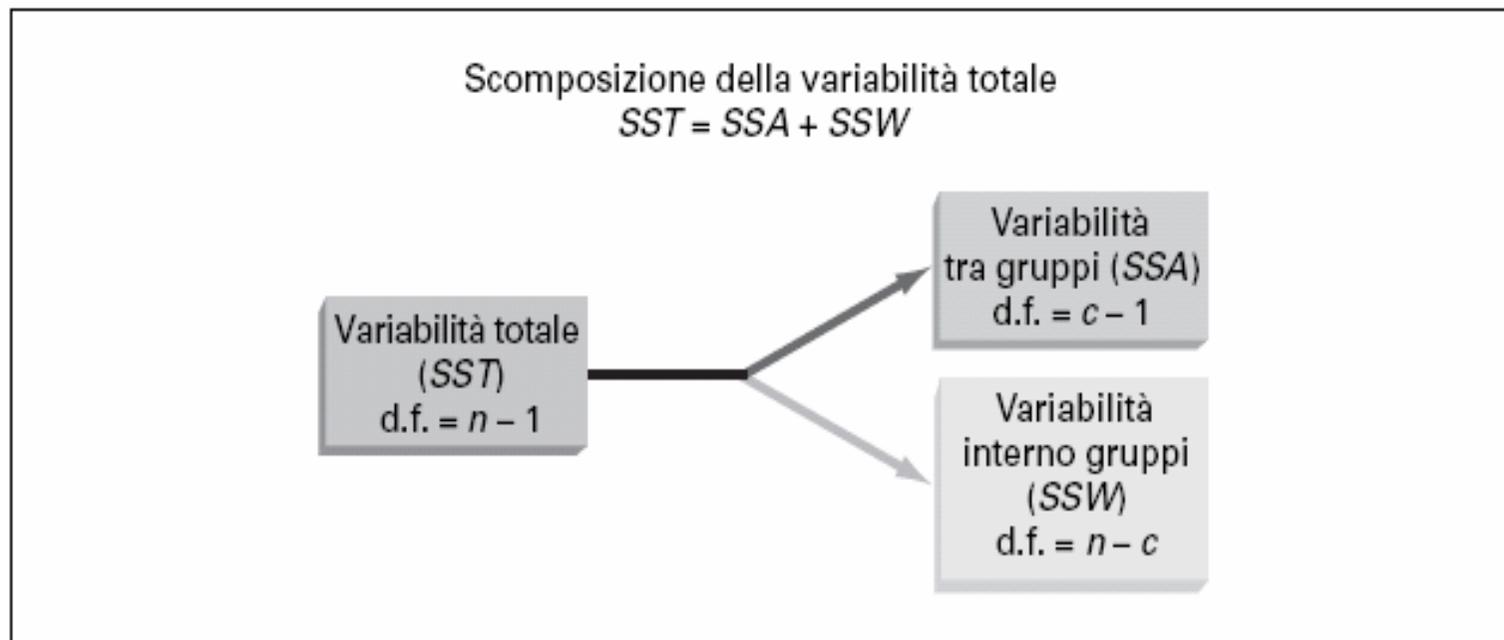
ANALISI DELLA VARIANZA

- **La variabilità all'interno dei gruppi è considerata un errore casuale, mentre la variabilità tra i gruppi è attribuibile alle differenze tra i gruppi, ed è anche chiamata effetto del trattamento**
- **Ipotizziamo che c gruppi rappresentino popolazioni con distribuzione normale, caratterizzate tutte dalla stessa varianza e che le osservazioni campionarie siano estratte casualmente ed indipendentemente dai c gruppi**
- **In questo contesto l'ipotesi nulla che si è interessati a verificare è che le medie di tutti i gruppi siano uguali tra loro, contro l'ipotesi alternativa che almeno una sia diversa**

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \dots = \mu_c \\ H_1 : \text{non tutte le medie sono uguali tra loro} \end{cases}$$

ANALISI DELLA VARIANZA

Per verificare le due ipotesi considerate, la variabilità totale (misurata dalla **somma dei quadrati totale – SST**) viene scomposta in due componenti: una componente attribuibile alla differenza tra i gruppi (misurata dalla **somma dei quadrati tra i gruppi – SSA**) e una seconda componente che si riferisce alle differenze riscontrate all'interno dei gruppi (misurata dalla **somma dei quadrati all'interno dei gruppi – SSW**)



ANALISI DELLA VARIANZA

$$SSA = \sum_{j=1}^c (\bar{x}_j - \bar{x})^2 n_j \text{ variabilità tra i gruppi}$$

$$SSW = \sum_{j=1}^c \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 \text{ variabilità entro i gruppi}$$

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

Dove:

c = numero dei gruppi

n_j = campione j - esimo

n = numerosità complessiva

j = generico gruppo

X_{ij} = generica osservazione appartenente al j - esimo gruppo

\bar{x}_j = media del j - esimo gruppo

\bar{x} = media generale

ANALISI DELLA VARIANZA

$(c-1)$ = gradi di libertà di SSW

$(n-c)$ = gradi di libertà di SSA

•Dividendo ciascuna somma dei quadrati per i rispettivi gradi di libertà, si ottengono tre varianze, o medie dei quadrati – **MSA** (la media dei quadrati tra gruppi), **MSW** (la media dei quadrati all'interno dei gruppi) e **MST** (la media dei quadrati totale).

•Se l'ipotesi nulla è vera e non ci sono differenze significative tra le medie dei gruppi, le tre medie dei quadrati – MSA, MSW e MST, che sono esse stesse delle stime di varianze e rappresentano tutte stime della varianza globale della popolazione sottostante

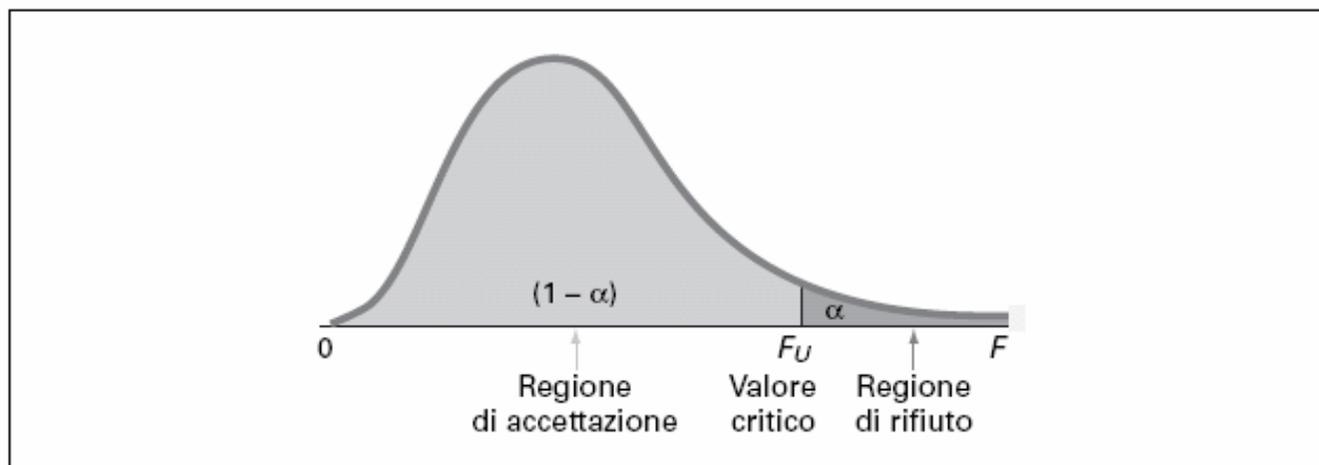
•Quindi per verificare l'ipotesi nulla contro l'alternativa si fa riferimento alla statistica test F per l'ANOVA a una via, ottenuta come rapporto tra MSA e MSW

ANALISI DELLA VARIANZA

Statistica test F per l'ANOVA ad una via

$$F = \frac{SSA / (n - c)}{SSW / (c - 1)} = \frac{MSA}{MSW}$$

- Se l'ipotesi nulla è vera, la realizzazione della statistica F dovrebbe essere approssimativamente 1, mentre se H_0 è falsa ci aspettiamo valori significatività superiori all'unità.
- La statistica F ha distribuzione F con $(c-1)$ gradi di libertà al numeratore e $(n-c)$ gradi di libertà al denominatore
- Quindi, fissato il livello di significatività α , l'ipotesi nulla dovrà essere rifiutata se il valore osservato della statistica test è maggiore del valore critico F_α di una distribuzione F con $(c-1)$ e $(n-c)$ gradi di libertà



ANALISI DELLA VARIANZA

I risultati del test F per l'ANOVA vengono solitamente riportati nella cosiddetta tabella dell'ANOVA

Fonte	Gradi di libertà	Somme dei quadrati	Medie dei quadrati (varianze)	F
Tra i gruppi	$c - 1$	SSA	$MSA = \frac{SSA}{c - 1}$	$F = \frac{MSA}{MSW}$
All'interno dei gruppi	$n - c$	SSW	$MSW = \frac{SSW}{n - c}$	

Nella tabella dell'ANOVA viene solitamente riportato anche il p-value, cioè la probabilità di osservare un valore di F maggiore o uguale a quello osservato, nel caso l'ipotesi nulla sia vera. Come usuale, l'ipotesi nulla di uguaglianza tra le medie dei gruppi deve essere rifiutata quando il p-value è inferiore al livello di significatività scelto.

ANALISI DELLA VARIANZA

esempio

Il responsabile delle vendite in una catena di supermercati vuole verificare se le vendite di un giocattolo per animali domestici possono essere influenzate dalla posizione del giocattolo sugli scaffali (posizione frontale, centrale, o posteriore). Considerato un campione di 18 punti vendita, le tre posizioni del giocattolo vengono sperimentate ciascuna in 6 punti vendita scelti casualmente. Nella tabella sono rappresentate le vendite del prodotto in ciascun punto vendita alla fine del periodo di prova.

FRONTALE	CENTRALE	POSTERIORE
8,6	3,2	4,6
7,2	2,4	6,0
5,4	2,0	4,0
6,2	1,4	2,8
5,0	1,8	2,2
4,0	1,6	2,8

- Ad un livello di significatività pari a 0,05 si può affermare che esiste una differenza significativa fra le vendite medie del prodotto ai diversi livelli del fattore;
- Quale scelta di locazione del prodotto sembra differire significativamente rispetto alle altre?
- Quali conclusioni dovrebbe trarre il responsabile alle vendite?

ANALISI DELLA VARIANZA

eempio

Il sistema d'ipotesi è

$$\begin{cases} H_0: \mu_1 = \mu_2 = \mu_3 \\ H_1: \text{almeno due medie sono diverse} \end{cases}$$

Preliminarmente calcoliamo le quantità seguenti:

$n_1 = 6$	$n_2 = 6$	$n_3 = 6$
$\bar{x}_1 = \sum x_{1i}/n_1 = 6,07$	$\bar{x}_2 = \sum x_{2i}/n_2 = 2,07$	$\bar{x}_3 = \sum x_{3i}/n_3 = 3,73$
$S_1^2 = \sqrt{\frac{\sum (x_{1i} - \bar{x}_1)^2}{n_1 - 1}} = 2,717$	$S_2^2 = \sqrt{\frac{\sum (x_{2i} - \bar{x}_2)^2}{n_2 - 1}} = 0,43$	$S_3^2 = \sqrt{\frac{\sum (x_{3i} - \bar{x}_3)^2}{n_3 - 1}} = 2,01$

Successivamente esponiamo i calcoli per il calcolo della ANOVA

Natura della variabilità	Devianza	Gradi di libertà	Varianze	F
Tra le classi	48,46	2	24,23	24,23:1,72 = 14,09
Entro le classi	25,75	15	1,72	
Totale	74,21	17	25,95	

ANALISI DELLA VARIANZA

esempio

Il risultato del test (14,09) deve essere confrontato con la v.c. F di Snedecor/Fisher in corrispondenza di 2 e 15 g.d.l. a livello di significatività del 5%. Il valore teorico è 3,68.

Il risultato del test cade nella zona di rifiuto che ci porta a concludere che:

a)vi è una differenza significativa tra le vendite medie del prodotto ai diversi livelli di fattori;

b)la scelta di locazione che sembra differire significativamente è la posizione frontale;

c)il direttore dovrebbe privilegiare il posizionamento centrale del prodotto.

ANALISI DELLA VARIANZA

esempio 2

*Scostamenti da un target specificato (in ml)
sotto 4 velocità di produzione*

	VELOCITÀ DI PRODUZIONE (IN BOTTIGLIE PER MINUTO)			
	210	240	270	300
	-3.5	3.4	-1.4	4.3
	2.0	-2.1	3.2	3.3
	-4.8	0.6	-1.2	2.0
	-2.1	-4.5	2.7	-0.8
	-4.0	-1.6	0.9	2.5
<i>Media</i>	$\bar{X}_1 = -2.48$	$\bar{X}_2 = -0.84$	$\bar{X}_3 = 0.84$	$\bar{X}_4 = 2.26$
<i>Varianza</i>	$S_1^2 = 7.237$	$S_2^2 = 8.903$	$S_3^2 = 4.553$	$S_4^2 = 3.683$
<i>Dev. st.</i>	$S_1 = 2.69$	$S_2 = 2.98$	$S_3 = 2.13$	$S_4 = 1.92$

Nota: uno scostamento negativo dal target denota una bottiglia "troppo vuota", mentre uno scostamento positivo dal target una bottiglia riempita eccessivamente (gli scostamenti sono misurati in millilitri).

ANALISI DELLA VARIANZA

esempio 2

Analisi della varianza per il problema di imbottigliamento industriale

Fonte	Gradi di Libertà	Somma dei Quadrati	Media dei Quadrati (Varianza)	F	p-value
Fra i gruppi (velocità)	4 - 1 = 3	63.286	21.095	3.46	0.041
Nei gruppi (velocità)	20 - 4 = 16	97.504	6.094		
Totale	20 - 1 = 19	160.790			

Al livello di significatività $\alpha=0,05$ rifiuto l'ipotesi di uguaglianza delle medie a favore dell'ipotesi che non tutte le medie sono uguali (p-value = 0,041 < 0,05). Il valore della distribuzione F con 3 e 16 g.l. che lascia a destra una probabilità di 0,05 è $F_u=3,24 < 3,46$.