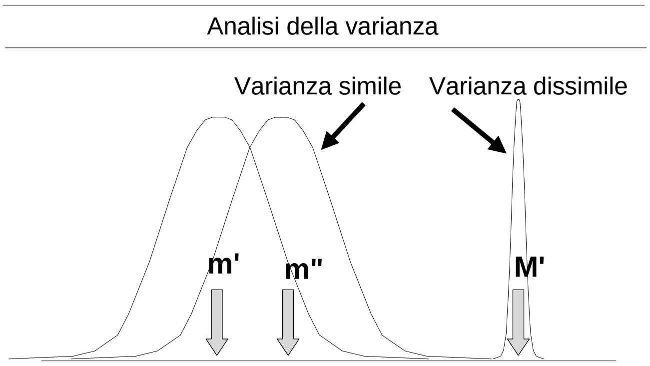
11. Analisi della varianza a un fattore

L’analisi della varianza generalizza il concetto del confronto tra medie, estendendolo al confronto contemporaneo tra più medie. Si considerino tre campioni con media rispettivamente m’, m” e M’. Le domande che ci si pone sono: la media m’ è significativamente diversa dalla media m”? E la media m’ è significativamente diversa dalla media M’? E la media m” è significativamente diversa dalla media M’?



A queste tre domande si può rispondere con l’analisi della varianza[1], che segnalerà che le tre distribuzioni presentano varianze dissimili (significativamente diverse) tra di loro, in questo caso essendo evidente che è la varianza che corrisponde alla distribuzione con media M’ ad essere la responsabile di tale differenza.

Se si parte dal presupposto che da una stessa popolazione sono estratti campioni con uguale varianza (e ovviamente uguale media, a meno di una differenza minima conseguente all’errore di campionamento), nel caso specifico si arriva a concludere, con un piccolo salto logico, che le medie campionarie sono significativamente diverse tra di loro. L’analisi della varianza è una tecnica generale, che si presta ad essere estesa a situazioni ancora più complesse, nelle quali peraltro alla complessità dei modelli adottati fa riscontro sempre la semplice base concettuale qui illustrata.

Sia i (con $i = 1,2,3,....r$) un generico campione, sia j (con $j = 1,2,3,....n$) un generico replicato, e quindi $x_{i,j}$ un generico valore della tabella

Campione	Replicato					Media
	j=1	j=2	j=3	. . . . .	j=n	$\bar{x}_i$
i=1	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	. . . . .	$x_{1,n}$	$\bar{x}_1$
i=2	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	. . . . .	$x_{2,n}$	$\bar{x}_2$
. . . .	. . . .	. . . .	. . . .	. . . . .	. . . .	. . . .
i=r	$x_{r,1}$	$x_{r,2}$	$x_{r,3}$	. . . . .	$x_{r,n}$	$\bar{x}_r$

nella quale le $r \cdot n$ osservazioni corrispondenti a r campioni, per ciascuno dei quali sono disponibili n dati, sono riportate in modo ordinato.

Siano ancora $\bar{x}_i$ la media di un generico campione, e $\bar{x}_g$ la media generale di tutte le $r \cdot n$ osservazioni. Allora la variabilità totale (S_t) osservata viene calcolata come

$$S_t = \sum_{i=1}^{i=r} \sum_{j=1}^{j=n} (x_{i,j} - \bar{x}_g)^2$$

[1] *Si rammenta che la varianza è il quadrato della deviazione standard, che a sua volta fornisce la misura dell’ampiezza della distribuzione.*

con $r \cdot n - 1$ gradi di libertà .

La variabilità S_s spiegata dalle differenza fra le medie $\bar{x}_i$ delle righe viene calcolata come

$$S_s = n \cdot \sum_{i=1}^{i=r} (\bar{x}_i - \bar{x}_g)^2$$

con $r - 1$ gradi di libertà, mentre la variabilità casuale, non spiegata, detta anche "residua" (S_n), viene calcolata come

$$S_n = \sum_{i=1}^{i=r} \sum_{j=1}^{j=n} (x_{i,j} - \bar{x}_i)^2$$

con $r \cdot (n - 1)$ gradi di libertà, tenendo presente che, per semplicità, essa può essere calcolata anche per differenza come

$$S_n = S_t - S_s$$

La varianza spiegata (V_s) e la varianza non spiegata (V_n), calcolate rispettivamente come

$$V_s = S_s / (r - 1)$$
$$V_n = S_n / (r \cdot (n - 1))$$

vengono allora impiegate per calcolare finalmente il rapporto fra varianze F

$$F = V_s / V_n$$

con $r - 1$ gradi di libertà al numeratore e $r \cdot (n - 1)$ gradi di libertà al denominatore.

Il valore di p corrispondente alla statistica F rappresenta la probabilità di osservare per caso un rapporto fra varianze della grandezza di quello effettivamente osservato: se tale probabilità è sufficientemente piccola, si conclude per una significatività della diversità fra le varianze, e conseguentemente che esistono delle differenze significative fra le medie $\bar{x}_i$ delle r righe.