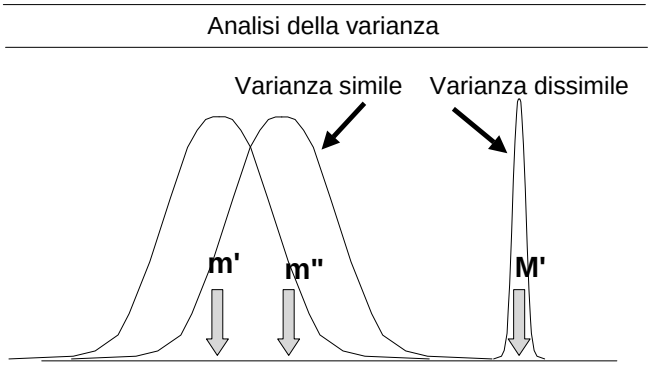
12. Analisi della varianza a due fattori

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.

Se l'analisi della varianza a 1 fattore (detta anche "a un criterio di classificazione") consente di verificare se vi siano (in media) differenze significative fra gli elementi appartenenti alle righe della tabella in cui sono stati ordinatamente raccolte le nostre osservazioni, l'ANOVA a 2 fattori consente di verificare contemporaneamente se vi siano (in media) differenze significative fra gli elementi appartenenti alle colonne della tabella, che si presenta così

Riga	Colonna					Media
	j=1	j=2	j=3	.....	j=c	$\bar{x}_{i.}$
i=1	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	.....	$x_{1,c}$	$\bar{x}_{1.}$
i=2	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	.....	$x_{2,c}$	$\bar{x}_{2.}$
....	....	....	....	.....	....	....
i=r	$x_{r,1}$	$x_{r,2}$	$x_{r,3}$	.....	$x_{r,c}$	$\bar{x}_{r.}$
Media $\bar{x}_{.j}$	$\bar{x}_{.1}$	$\bar{x}_{.2}$	$\bar{x}_{.3}$	.....	$\bar{x}_{.c}$	

Ciascuno dei dati $x_{i,j}$ risulta pertanto assegnato in base a una caratteristica a una delle i ($i = 1,2,3,....r$) righe e per un'altra caratteristica a una delle j ($j = 1,2,3,....c$) colonne.

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

Siano allora $\bar{x}_{i.}$ la media di una generica riga, $\bar{x}_{.j}$ la media di una generica colonna, e $\bar{x}_g$ la media generale degli $r \cdot c$ dati. È facile notare che, rispetto alla tabella dell'analisi della varianza a 1 fattore, la novità consiste nell'avere introdotto un elemento di classificazione a livello delle colonne, e quindi le corrispondenti medie $\bar{x}_{.j}$.

La variabilità totale (S_t) osservata viene calcolata come

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

con $r \cdot c - 1$ gradi di libertà.

Tuttavia, essendo stato introdotto un nuovo criterio di classificazione dei dati, essa verrà scomposta non più in due, bensì in tre componenti. La prima di esse, la variabilità S_r spiegata dalle differenze fra le medie $\bar{x}_{i.}$, cioè dalle differenze fra le medie delle righe, viene calcolata come

$$S_r = c \cdot \sum_{i=1}^{i=r} (\bar{x}_{i.} - \bar{x}_g)^2$$

con $r - 1$ gradi di libertà.

La variabilità S_c spiegata dalle differenze fra le medie $\bar{x}_{.j}$, cioè dalle differenze fra le medie delle colonne, viene calcolata come

$$S_c = r \cdot \sum_{j=1}^{j=c} (\bar{x}_{.j} - \bar{x}_g)^2$$

con $c - 1$ gradi di libertà.

Infine la variabilità casuale, non spiegata (S_n), detta anche "residua", viene calcolata come

$$S_n = \sum_{i=1}^{i=r} \sum_{j=1}^{j=c} (x_{i,j} - \bar{x}_{i.} - \bar{x}_{.j} - \bar{x}_g)^2$$

con $(r - 1) \cdot (c - 1)$ gradi di libertà, tenendo presente che può più semplicemente essere calcolata per differenza come

$$S_n = S_t - S_r - S_c$$

La varianza spiegata dalle differenze fra le medie $\bar{x}_{i.}$ delle righe (V_r), quella spiegata dalle differenze fra le medie $x_{.j}$ delle colonne (V_c) e la varianza non spiegata (V_n) sono allora calcolate rispettivamente come

$$V_r = S_r / (r - 1)$$
$$V_c = S_c / (c - 1)$$
$$V_n = S_n / ((r - 1) \cdot (c - 1))$$

Il rapporto fra varianze F

$$F = V_r / V_n$$

con $r - 1$ gradi di libertà al numeratore e $(r - 1) \cdot (c - 1)$ gradi di libertà al denominatore viene allora impiegato per verificare l'esistenza di una differenza significativa fra le medie delle righe ($\bar{x}_{i.}$).

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.

Il rapporto fra varianze

$$F = V_c / V_n$$

con $c - 1$ gradi di libertà al numeratore e $(r - 1) \cdot (c - 1)$ gradi di libertà al denominatore viene impiegato per verificare l'esistenza di una differenza significativa fra le medie delle colonne ($\bar{x}_{.j}$).

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 $\bar{x}_j$ delle c colonne.