ANALISI DELLA VARIANZA

TEMI CONSIDERATI:

1. **Confronto simultaneo di più valori medi**

2. **Scomposizione della varianza di z=x+y**

3. **Distribuzione del rapporto tra varianze**

4. **Test di omogeneità per k medie:**
   - **stima di σ^2 con la varianza *intra*classe**
   - **stima di σ^2 con la varianza *inter*classe**

ANALISI DELLA VARIANZA:

Si abbiano k campioni. Vogliamo vedere se si possono considerare provenienti da insiemi che hanno lo stesso valore centrale (media) o la stessa dispersione (varianza).

In questo caso, l'ipotesi da verificare è l'*omogeneità delle k variabili aleatorie*, cioè le k medie oppure le k varianze dei campioni.

Se tutti i confronti fossero prossimi alla soglia di significatività, si potrebbe concludere che vi è eterogeneità degli insiemi considerati, ma ciò non sarebbe messo in evidenza in maniera obiettiva.

Con un gran numero di campioni, se si confrontano le caratteristiche due a due al livello di significatività P=0,05, in media il 5% delle differenze saranno significative, anche se in realtà gli insiemi sono omogenei e le poche differenze significative potrebbero essere dovute al caso e non a differenze reali tra gli insiemi.

Occorrono, quindi, metodi che siano più adatti a verificare l'ipotesi di omogeneità di un insieme di medie e di un insieme di varianze.

Prima di procedere a questo esame, conviene mostrare anzitutto come si possono *scomporre la media e la varianza generale di un insieme formato da più sottoinsiemi di dati*, in modo da mettere in evidenza le relazioni tra le statistiche dell'insieme e quelle corrispondenti dei suoi sottoinsiemi.

Scomposizione della varianza:

Le varianze hanno la proprietà di essere spesso additive, mentre *la somma dei quadrati degli scarti è sempre additiva*. Per contro la deviazione standard non ha questo vantaggio.

Siano x ed y due variabili indipendenti e sia z = x + y la variabile somma delle due variabili date. Poiché è:

$$\mu_z = \mu_x + \mu_y$$

In generale si ha:

$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2 + 2\sigma_{xy}$$

Se le variabili sono indipendenti, si potrà scrivere l'uguaglianza:

$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2$$

Nell'*ANOVA* Si separa la variazione totale nelle sue varie componenti cercando quale proporzione di essa può essere attribuita alla varianza di ciascun gruppo.

Si abbiano k campioni per un totale di $n=\Sigma n_i$ osservazioni così distribuite:

1	**2**		**i**		**k**
x_{11}	x_{21}	....	x_{i1}	....	x_{k1}
x_{12}	x_{22}	....	x_{i2}	....	x_{k2}
x_{13}	x_{23}	....	x_{i3}	....	x_{k3}
....	....	....	....	....	....
x_{1j}	x_{2j}	...	x_{ij}	...	x_{kj}
....	....	....	....	....	....
x_{1n_1}	x_{2n_2}	...	x_{in_i}	...	x_{kn_k}
Frequenze n_1	n_2	....	n_i	....	n_k

Media generale: $\overline{\overline{X}}_{**}$

Medie: $\overline{X}_{1*}$ $\overline{X}_{2*}$ $\overline{X}_{k*}$

Scarti totali rispetto alla media

$$\sum_{i=1}^{k}\sum_{j=1}^{nk}\left(x_{ij}-\overline{\overline{X}}_{**}\right)^2=\sum_{i=1}^{k}\sum_{j=1}^{nk}\left(x_{ij}-\overline{X}_{i*}+\overline{X}_{i*}-\overline{\overline{X}}_{**}\right)^2=$$

$$=\sum_{i=1}^{k}\sum_{j=1}^{nk}\left(x_{ij}-\overline{X}_{i*}\right)^2+\sum_{i=1}^{k}\sum_{j=1}^{nk}\left(\overline{X}_{i*}-\overline{\overline{X}}_{**}\right)^2+2\sum_{i=1}^{k}\sum_{j=1}^{nk}\left(x_{ij}-\overline{X}_{i*}\right)\left(\overline{X}_{i*}-\overline{\overline{X}}_{**}\right)=$$

$$=\sum_{i=1}^{k}\sum_{j=1}^{nk}\left(x_{ij}-\overline{X}_{i*}\right)^2+\sum_{i=1}^{k}\sum_{j=1}^{nk}\left(\overline{X}_{i*}-\overline{\overline{X}}_{**}\right)^2+2\sum_{i=1}^{k}\left[\left(\overline{X}_{i*}-\overline{\overline{X}}_{**}\right)\sum_{j=1}^{nk}\left(x_{ij}-\overline{X}_{i*}\right)\right]=$$

$$=\sum_{i=1}^{k}\sum_{j=1}^{nk}\left(x_{ij}-\overline{X}_{i*}\right)^2+\sum_{i=1}^{k}\sum_{j=1}^{nk}\left(\overline{X}_{i*}-\overline{\overline{X}}_{**}\right)^2+2\sum_{i=1}^{k}\left[\left(\overline{X}_{i*}-\overline{\overline{X}}_{**}\right)\sum_{j=1}^{nk}\left(x_{ij}-\overline{X}_{i*}\right)\right]=$$

$$=\sum_{i=1}^{k}\sum_{j=1}^{nk}\left(x_{ij}-\overline{X}_{i*}\right)^2+\sum_{i=1}^{k}\sum_{j=1}^{nk}\left(\overline{X}_{i*}-\overline{\overline{X}}_{**}\right)^2+2\sum_{i=1}^{k}\left[\left(\overline{X}_{i*}-\overline{\overline{X}}_{**}\right)\left(\sum_{j=1}^{nk}x_{ij}-nk\overline{X}_{i*}\right)\right]=$$

$$=\sum_{i=1}^{k}\sum_{j=1}^{nk}\left(x_{ij}-\overline{X}_{i*}\right)^2+\sum_{i=1}^{k}\sum_{j=1}^{nk}\left(\overline{X}_{i*}-\overline{\overline{X}}_{**}\right)^2$$

Se nk=n si ha:

$$\sum_{i=1}^{k}\sum_{j=1}^{nk}\left(x_{ij}-\overline{\overline{X}}_{**}\right)^2=\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij}-\overline{X}_{i*}\right)^2+\sum_{i=1}^{k}\sum_{j=1}^{n}\left(\overline{X}_{i*}-\overline{\overline{X}}_{**}\right)^2=$$

$$=\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij}-\overline{X}_{i*}\right)^2+n\sum_{i=1}^{k}\left(\overline{X}_{i*}-\overline{\overline{X}}_{**}\right)^2$$

cioè:

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij}-\overline{\overline{X}}_{**}\right)^2}_{\text{Somma degli scarti quadratici totali}}=\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij}-\overline{X}_{i*}\right)^2}_{\text{Somma degli scarti quadratici dei gruppi}}+\underbrace{n\sum_{i=1}^{k}\left(\overline{X}_{i*}-\overline{\overline{X}}_{**}\right)^2}_{\text{Somma degli scarti quadratici delle medie}}$$

Somma degli scarti quadratici totali

$$\underbrace{\frac{\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij}-\overline{\overline{X}}_{**}\right)^2}{\sigma^2}}_{\text{Somma degli scarti quadratici totali}}=\underbrace{\frac{\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij}-\overline{X}_{i*}\right)^2}{\sigma^2}}_{\text{Somma degli scarti quadratici dei gruppi}}+\underbrace{\frac{n\sum_{i=1}^{k}\left(\overline{X}_{i*}-\overline{\overline{X}}_{**}\right)^2}{\sigma^2}}_{\text{Somma degli scarti quadratici delle medie}}$$

$\chi^2_{\text{kn-1}}$ $\chi^2_{\text{k(n-1)}}$ $\chi^2_{\text{k-1}}$

Somma degli scarti quadratici dei gruppi
Stima di σ^2 con la varianza *intra*classe

$$S_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_{ij} - \overline{X}_{i*})^2$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{k}(n-1)S_i^2}{\sum_{i=1}^{k}(n-1)} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij} - \overline{X}_{i*}\right)^2}{k(n-1)}$$

$$\frac{\hat{\sigma}^2}{\sigma^2} = \frac{\sum_{i=1}^{k}\frac{\sum_{j=1}^{n}\left(x_{ij} - \overline{X}_{i*}\right)^2}{\sigma^2}}{k(n-1)} \simeq \frac{\sum_{i=1}^{k}\chi^2_{n-1}}{k(n-1)} = \frac{\chi^2_{k(n-1)}}{k(n-1)}$$

Somma degli scarti quadratici delle medie
Stima di σ^2 con la varianza *inter*classe

$$S^2_{\overline{X}} = \frac{\sum_{i=1}^{k}\left(\overline{X}_{i*} - \overline{\overline{X}}_{**}\right)^2}{k-1}$$

se vero: $H_0 : \mu_1 = \mu_2 = ... = \mu_K = \mu$

$$\frac{(k-1)nS^2_{\overline{X}}}{\sigma^2} = \frac{(k-1)\sum_{i=1}^{k}\left(\overline{X}_{i*} - \overline{\overline{X}}_{**}\right)^2}{(k-1)\sigma^2/n} = \frac{n\sum_{i=1}^{k}\left(\overline{X}_{i*} - \overline{\overline{X}}_{**}\right)^2}{\sigma^2} \simeq \chi^2_{k-1}$$

TEST SULL'EGUAGLIANZA TRA PIÙ MEDIE, OVVERO ANALISI DELLA VARIANZA AD UNA VIA

Si verifica l'effetto di un criterio di raggruppamento sulla distribuzione del campione. La generica realizzazione della v.a. x_{ij}, con *i* il gruppo di appartenenza, è esprimibile come:

$$x_{ij} = \mu + \delta_j + \varepsilon_{ij} \qquad \forall i = 1,...,n_k;\ j = 1,...,K$$

$$\sum_{k=1}^{K} n_k = N; \quad \delta_j = \mu_j - \mu; \quad \varepsilon_{ij} = x_{ij} - \mu_j$$

δ_k è l'errore sistematico (o effetto) dovuto al gruppo
ε_{ik} è l'errore casuale di campionamento

$$H_0: \mu_1 = \mu_2 = ... = \mu_K = \mu \ \Leftrightarrow H_0: \delta_1 = \delta_2 = ... = \delta_K = 0$$

Se vi è un errore sistematico di gruppo δ_k, la varianza totale sarà la somma della varianza di gruppo e di quella di campionamento. Testando la differenza tra la varianza di gruppo e quella di campionamento si verifica se appartengono o meno alla stessa popolazione ossia sono entrambi errori di campionamento.

v.a.	Somma scarti quad.	g.d.l.	
ε_{ij}	$SSA = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij} - \overline{X}_{i*}\right)^2$	$k(n-1)$	$F = \frac{SSA}{SSE}\frac{k-1}{k(n-1)}$
δ_k	SSE= $n\sum_{i=1}^{k}\left(\overline{X}_{i*} - \overline{\overline{X}}_{**}\right)^2$	k-1	
$x_{ik} - \mu_k$	$SST = SSA + SSE = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij} - \overline{\overline{X}}_{**}\right)^2$	nk-1	

Si effettua un test unilaterale F

Camp.1	Camp.2	Camp.3
28.14407	28.0383	15.33615
14.50595	37.67061	16.53886
14.74611	33.63762	24.1739
14.93495	25.92654	35.86248
18.42531	32.86984	22.55039
12.92439	32.37778	24.41302
17.56323	33.23385	12.89464
14.44797	29.77826	24.8458
18.61809	39.54465	13.33999
15.75955	37.76038	24.01915

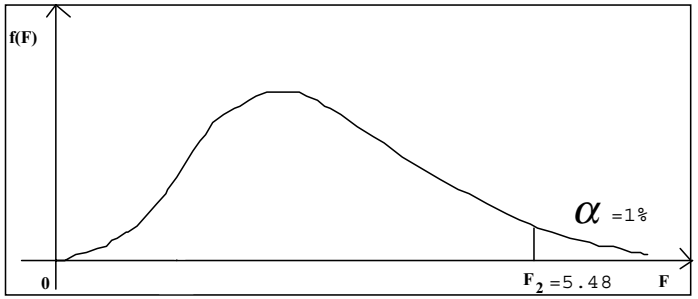
ESEMPIO
Popolazioni normali con media 20,30,25 e σ=6

$\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij}-\overline{X}_{i*}\right)^2$=786.2179241

$n\sum_{i=1}^{k}\left(\overline{X}_{i*}-\overline{\overline{X}}_{**}\right)^2$=1381.037

$\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij}-\overline{\overline{X}}_{**}\right)^2=2167.254564$

$$F=\frac{\dfrac{n\sum_{i=1}^{k}\left(\overline{X}_{i*}-\overline{\overline{X}}_{**}\right)^2}{k-1}}{\dfrac{\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij}-\overline{X}_{i*}\right)^2}{k(n-1)}}=\frac{\dfrac{1381.03664}{2}}{\dfrac{786.2179241}{27}}=23.71$$



Camp.1	Camp.2	Camp.3
28.14407	28.0383	15.33615
14.50595	37.67061	16.53886
14.74611	33.63762	24.1739
14.93495	25.92654	35.86248

ESEMPIO
Popolazioni normali con media 20,30,25 e σ=6

$\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij}-\overline{X}_{i*}\right)^2$=487.92

$n\sum_{i=1}^{k}\left(\overline{X}_{i*}-\overline{\overline{X}}_{**}\right)^2$=358.27

$\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij}-\overline{\overline{X}}_{**}\right)^2=846.1967542$

scarti q risp media

101.2298	10.75817	58.39557
12.79365	40.35232	41.46051
11.1333	5.379375	1.430539
9.90878	29.07075	166.0137
Somma=488		

	x1*	x2*	x2*	
	18.08277	31.31827	22.97785	
x**	24.1263			

Scarti quadrati rispetto media

36.52416	51.72446	1.318937
Somma*4	358.27	

$$F = \frac{\dfrac{n\sum_{i=1}^{k}\left(\overline{X}_{i*} - \overline{\overline{X}}_{**}\right)^2}{k-1}}{\dfrac{\sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij} - \overline{X}_{i*}\right)^2}{k(n-1)}} = \frac{\dfrac{358.27}{2}}{\dfrac{487.92}{9}} = 3.30$$