L'Analisi della Varianza (ANOVA)

Lucio Demeio
Dipartimento di Ingegneria Industriale e Scienze Matematiche
Università Politecnica delle Marche

Introduzione. L'analisi della varianza (indicata spesso con l'acronimo ANOVA, dalla terminologia inglese *ANalysis Of VAriance*) è un test d'ipotesi volto a confermare (o respingere) l'affermazione che m campioni gaussiani, non necessariamente aventi lo stesso rango, presi da popolazioni diverse ma tutte con la stessa varianza σ^2, abbiano tutti lo stesso valor medio. Trattiamo qui soltanto l'Analisi della Varianza *ad una via*, tralasciando la più complessa Analisi della Varianza *a due vie*.

$$
\begin{aligned}
&X_{11}, X_{12}, ..., X_{1,n_1}\\
&X_{21}, X_{22}, ..., X_{2,n_2}\\
&...\\
&X_{m1}, X_{m2}, ..., X_{m,n_m}
\end{aligned}
$$

gli m campioni, di rango rispettivamente $n_1, n_2, ..., n_m$. Siano inoltre $\mu_1, \mu_2, ..., \mu_m$ i valori medi delle variabili dei singoli campioni, σ^2 la varianza comune e siano $\overline{X}_i$, $i = 1, 2, ..., m$ le medie campionarie. Siano infine μ la media totale $\overline{X}_{tot}$ la media campionaria totale. Abbiamo:

$$
\begin{aligned}
&X_{ij} \sim N(\mu_i, \sigma^2) \qquad \forall\ i = 1, 2, ..., m, \quad j = 1, 2, ..., n_i\\
&\overline{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} \qquad \forall\ i = 1, 2, ..., m\\
&\overline{X}_i \sim N(\mu_i, \sigma^2/n_i) \qquad \forall\ i = 1, 2, ..., m\\
&\mu = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n_i} E[X_{ij}] = \sum_{i=1}^{m} \frac{n_i}{N}\, \mu_i\\
&\overline{X}_{tot} = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n_i} X_{ij} = \sum_{i=1}^{m} \frac{n_i}{N}\, \overline{X}_i\\
&\overline{X}_{tot} \sim N(\mu, \sigma^2/N)
\end{aligned}
$$

dove $N = n_1 + n_2 + ... + n_m$. Con queste premesse, l'ipotesi nulla e l'ipotesi alternativa sono espresse da

$$
\begin{aligned}
&H_0 : \mu_1 = \mu_2 = ... = \mu_m = \mu\\
&H_1 : \mu_{i_1} \neq \mu_{i_2} \qquad \text{per almeno una coppia} \quad i_1, i_2.
\end{aligned}
$$

Il test consiste nel costruire due diversi stimatori della varianza, $\widehat{\sigma^2}_W$ e $\widehat{\sigma^2}_b$, con $\widehat{\sigma^2}_W$ uno stimatore corretto sia con H_0 vera che con H_0 falsa, mentre $\widehat{\sigma^2}_b$ ha distorsione non nulla nel

caso di H_0 falsa. Si dimostra quindi che

$$E[\widehat{\sigma^2}_b] \geq \sigma^2$$

con il segno di uguaglianza valido solo nel caso di H_0 vera. La statistica da usare per la verifica del test è pertanto

$$F \equiv \frac{\widehat{\sigma^2}_b}{\widehat{\sigma^2}_W}. \tag{1}$$

La regione critica di livello α è perciò data da

$$P(F > F_\alpha) = \alpha. \tag{2}$$

Dobbiamo pertanto:

- costruire gli stimatori $\widehat{\sigma^2}_W$ e $\widehat{\sigma^2}_b$;
- determinare la legge del loro rapporto $\widehat{\sigma^2}_b/\widehat{\sigma^2}_W$;
- costruire la regione critica del test.

La variazione totale e la formula di Huygens. Introduciamo la *variazione totale delle osservazioni* S^2 definita come

$$S^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{tot})^2. \tag{3}$$

Questa grandezza si può decomporre al modo seguente:

$$\begin{aligned} S^2 &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{tot})^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i + \overline{X}_i - \overline{X}_{tot})^2 = \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left\{(X_{ij} - \overline{X}_i)^2 + (\overline{X}_i - \overline{X}_{tot})^2 + 2\,(X_{ij} - \overline{X}_i)\,(\overline{X}_i - \overline{X}_{tot})\right\} = \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2 + \sum_{i=1}^{m} n_i\,(\overline{X}_i - \overline{X}_{tot})^2 = S_W^2 + S_b^2 \end{aligned} \tag{4}$$

dove abbiamo usato il fatto che

$$\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i) = 0$$

e dove abbiamo introdotto le grandezze

$$S_W^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2 \tag{5}$$

$$S_b^2 = \sum_{i=1}^{m} n_i\,(\overline{X}_i - \overline{X}_{tot})^2 \tag{6}$$

La (4) dice che la variazione totale è data dalla somma delle variazioni di ogni gruppo dalla sua media campionaria (S_W^2) e delle variazioni tra i gruppi (S_b^2). Il contributo S_W^2 viene anche detto variazione *entro (within)* i campioni, mentre S_b^2 viene detto variazione *tra (between)* i campioni. La (4) è detta *formula di Huygens* ed ha la stessa struttura matematica del teorema di Huygens nella meccanica dei corpi rigidi.

Gli stimatori $\widehat{\sigma^2}_W$ e $\widehat{\sigma^2}_b$. All'interno di ciascun campione, ad esempio per la variabile X_{ij} del campione $i-$esimo, abbiamo

$$Z_{ij} = \frac{X_{ij} - \mu_i}{\sigma} \sim N(0,1) \qquad \forall j = 1, 2, ..., n_i$$

e quindi

$$\sum_{j=1}^{n_i} Z_{ij}^2 \sim \chi_{n_i}^2 \quad \text{e} \quad \sum_{i=1}^{m}\sum_{j=1}^{n_i} Z_{ij}^2 \sim \chi_N^2$$

Sappiamo che, sostituendo le medie campionarie $\overline{X}_i$ al posto delle medie vere μ_i,

$$\sum_{j=1}^{n_i} \frac{(X_{ij} - \overline{X}_i)^2}{\sigma^2} \sim \chi_{n_i-1}^2 \quad \text{e} \quad \sum_{i=1}^{m}\sum_{j=1}^{n_i} \frac{(X_{ij} - \overline{X}_i)^2}{\sigma^2} \sim \chi_{N-m}^2$$

e quindi, ricordando la proprietà della legge χ^2 per cui $E[\chi_l^2] = l$, abbiamo per la grandezza S_W^2 introdotta nella (5)

$$E\left[\frac{S_W^2}{N-m}\right] = \sigma^2.$$

Lo stimatore

$$\widehat{\sigma^2}_W = \frac{S_W^2}{N-m} \tag{7}$$

è pertanto uno stimatore non distorto della varianza σ^2. Siccome $\widehat{\sigma^2}_W$ è stato introdotto senza l'ipotesi di uguaglianza delle medie dei singoli campioni, esso è uno stimatore corretto di σ^2 indipendentemente dalla validità o meno dell'ipotesi nulla H_0.

Per determinare il secondo stimatore, $\widehat{\sigma^2}_b$, consideriamo la grandezza S_b^2 introdotta nella (6). Abbiamo:

$$\begin{aligned}
S_b^2 &= \sum_{i=1}^{m} n_i \left(\overline{X}_i - \overline{X}_{tot}\right)^2 = \sum_{i=1}^{m} n_i \left\{\overline{X}_i - \mu_i + \mu - \overline{X}_{tot} + \mu_i - \mu\right\}^2 = \\
&= \sum_{i=1}^{m} n_i \left\{(\overline{X}_i - \overline{X}_{tot} - \mu_i + \mu)^2 + (\mu_i - \mu)^2 + 2\left(\mu_i - \mu\right)\left(\overline{X}_i - \overline{X}_{tot} - \mu_i + \mu\right)\right\} = \\
&= \sum_{i=1}^{m} n_i \left(\overline{X}_i - \overline{X}_{tot} - \mu_i + \mu\right)^2 + \sum_{i=1}^{m} n_i \left(\mu_i - \mu\right)^2 + \\
&\quad +2 \sum_{i=1}^{m} n_i \left(\mu_i - \mu\right)\left(\overline{X}_i - \overline{X}_{tot} - \mu_i + \mu\right)
\end{aligned}$$

Notiamo che

$$\begin{aligned}
&E[\overline{X}_i - \overline{X}_{tot} - \mu_i + \mu] = 0 \\
&E[(\overline{X}_i - \overline{X}_{tot} - \mu_i + \mu)^2] = Var(\overline{X}_i - \overline{X}_{tot}) \\
&E[(\mu_i - \mu)^2] = (\mu_i - \mu)^2 \geq 0;
\end{aligned}$$

e pertanto

$$E[S_b^2] = \sum_{i=1}^{m} n_i \, Var(\overline{X}_i - \overline{X}_{tot}) + \sum_{i=1}^{m} n_i \, (\mu_i - \mu)^2.$$

Per la prima sommatoria abbiamo:

$$\begin{aligned}
\sum_{i=1}^{m} n_i \, Var(\overline{X}_i - \overline{X}_{tot}) &= \sum_{i=1}^{m} n_i \, \left\{ Var(\overline{X}_i) + Var(\overline{X}_{tot}) - 2\, Cov(\overline{X}_i, \overline{X}_{tot}) \right\} = \\
&= \sum_{i=1}^{m} n_i \left\{ \frac{\sigma^2}{n_i} + \frac{\sigma^2}{N} - 2\, Cov\left(\overline{X}_i, \sum_{k=1}^{m} \frac{n_k}{N} \overline{X}_k \right) \right\} = \\
&= \sum_{i=1}^{m} n_i \left\{ \frac{\sigma^2}{n_i} + \frac{\sigma^2}{N} - 2 \sum_{k=1}^{m} \frac{n_k}{N} \, Cov\left(\overline{X}_i, \overline{X}_k \right) \right\} = \\
&= \sum_{i=1}^{m} n_i \left\{ \frac{\sigma^2}{n_i} + \frac{\sigma^2}{N} - 2 \, \frac{n_i}{N} \, Var(\overline{X}_i) \right\} = \sum_{i=1}^{m} n_i \left\{ \frac{\sigma^2}{n_i} + \frac{\sigma^2}{N} - 2 \, \frac{n_i}{N} \, \frac{\sigma^2}{n_i} \right\} = \\
&= \sigma^2 \sum_{i=1}^{m} n_i \left(\frac{1}{n_i} - \frac{1}{N} \right) = (m-1)\, \sigma^2
\end{aligned}$$

In conclusione,

$$E[S_b^2] = (m-1)\, \sigma^2 + \sum_{i=1}^{m} n_i \, (\mu_i - \mu)^2$$

ovvero

$$E\left[\frac{S_b^2}{m-1} \right] = \sigma^2 + \frac{1}{m-1} \sum_{i=1}^{m} n_i \, (\mu_i - \mu)^2 \tag{8}$$

Dalla (8) vediamo che

$$\widehat{\sigma^2}_b = \frac{S_b^2}{m-1} \tag{9}$$

è uno stimatore non distorto della varianza se e solo se è vera l'ipotesi nulla H_0. Inoltre, in generale a prescindere dalla validità o meno dell'ipotesi nulla, si ha che

$$E[\widehat{\sigma^2}_b] \geq E[\widehat{\sigma^2}_W] = \sigma^2. \tag{10}$$

La distribuzione di Snedecor. La statistica F introdotta nella (1) è quindi data da

$$F = F_{m-1,N-m} \equiv \frac{S_b^2}{S_W^2} \, \frac{N-m}{m-1}. \tag{11}$$

La distribuzione di questa variabile viene detta *distribuzione F di Snedecor (o Fisher-Snedecor)* ed è definita come il rapporto tra due distribuzioni χ^2 al modo seguente:

$$F_{kl} = \frac{\chi_k^2 / k}{\chi_l^2 / l}.$$

I quantili $F_{\alpha,kl}$ della legge di Snedecor sono definiti in maniera analoga ai quantili della normale standard o della legge di Student. Quindi avremo

$$P(F_{kl} > F_\alpha(k,l)) = \alpha.$$

In base alla disuguaglianza (10), definiamo pertanto la regione di rigetto dell'ipotesi nulla al livello di confidenza α, come

$$P(F_{m-1,N-m} > F_\alpha(m-1, N-m)) = \alpha. \tag{12}$$

Esercizio (Ross, Cap. 10 n. 6) Per confrontare l'efficacia di due diete si scelgono 20 individui sovrappeso e li si divide a caso in due gruppi da 10, ciascuno dei quali viene sottoposto ad una delle due diete. Dopo 10 settimane le diminuzioni di peso riscontrate sono state:

Dieta 1:	$22.2, 23.4, 24.2, 16.1, 9.4, 12.5, 18.6, 32.2, 8.8, 7.6$
Dieta 2:	$24.2, 16.8, 14.6, 13.7, 19.5, 17.6, 11.2, 9.5, 30.1, 21.5$

Verifica al 5% di significatività che le due diete abbiano avuto lo stesso effetto.

Soluzione. I due campioni hanno lo stesso rango, pari a 10. Quindi in questo caso $m = 2$ e $n_1 = n_2 = n = 10$, $N = 20$. Indichiamo con X_{11}, X_{12}, ..., X_{1n} gli elementi del primo campione e con X_{21}, X_{22}, ..., X_{2n} quelli del secondo. Dai dati abbiamo:

$$\begin{aligned}
&\overline{X}_1 = 17.5, \qquad \overline{X}_2 = 17.87, \qquad \overline{X}_{tot} = 17.685\\
&S_W^2 = 934.68, \qquad S_b^2 = 0.68\\
&\widehat{\sigma^2}_W = 51.93, \qquad \widehat{\sigma^2}_b = 0.68\\
&F_{1,18} = \frac{\widehat{\sigma^2}_b}{\widehat{\sigma^2}_W} = 0.013.
\end{aligned}$$

Il quantile di ordine 0.05 che rileviamo dalle tavole è $F_{0.05}(1, 18) = 4.41$; l'ipotesi nulla non può dunque essere rifiutata.

Esercizio (Baldi, Cap. 6 Esempio 6.48) Tre tipi diversi di terreno vengono coltivati ad orzo; per ogni tipo di terreno si misura il raccolto in quintali per ettaro; le quantità prodotte per ogni tipo di terreno sono:

Tipo 1:	$40.3, 52.1, 46.5, 46.5, 52.1, 48.3, 45.3, 45.1, 41.8, 39.8, 47.0$
Tipo 2:	$44.5, 51.0, 42.5, 49.3, 45.7, 46.0, 54.8, 39.4, 51.1, 45.6, 56.0, 52.2, 47.2$
Tipo 3:	$44.1, 48.5, 41.3, 42.6, 42.1, 43.8, 39.7$

Verifica al 5% di significatività che le due diete abbiano avuto lo stesso effetto.

Soluzione. I tre campioni hanno rango, rispettivamente 13, 11 e 7. Quindi in questo caso $m = 3$ e $n_1 = 13, n_2 = 11, n_3 = 7$, $N = 31$. Indichiamo con X_{11}, X_{12}, ..., X_{1n_1} gli elementi del primo campione, con X_{21}, X_{22}, ..., X_{2n_2} quelli del secondo e con X_{31}, X_{32}, ..., X_{3n_3}. Dai dati abbiamo:

$$\begin{aligned}
&\overline{X}_1 = 45.89, \qquad \overline{X}_2 = 48.1, \qquad \overline{X}_3 = 43.16, \overline{X}_{tot} = 46.2\\
&S_W^2 = 497.83, \qquad S_b^2 = 112.79\\
&\widehat{\sigma^2}_W = 17.17, \qquad \widehat{\sigma^2}_b = 56.40\\
&F_{1,18} = \frac{\widehat{\sigma^2}_b}{\widehat{\sigma^2}_W} = 3.29.
\end{aligned}$$

Il quantile di ordine 0.05 che rileviamo dalle tavole è $F_{0.05}(2, 28) = 3.34$; l'ipotesi nulla non può dunque essere rifiutata.