

12 Analisi della varianza (cenni)*

12.1 Introduzione

L'analisi della varianza o **ANOVA** (inventata da R.A. Fisher) è una metodologia generale per fare inferenze su un insieme di parametri relativi a medie di popolazioni. L'ANOVA può essere ad uno o a due fattori. Si suppone che i dati siano estratti da popolazioni normali con la stessa varianza σ^2 , che non è nota. Il metodo dell'ANOVA per verificare un'ipotesi nulla H_0 riguardante più parametri, richiede di derivare due stimatori della varianza comune σ^2 .

Il primo stimatore T_1 di σ^2 è valido sia che l'ipotesi nulla H_0 sia vera o sia falsa. Il secondo stimatore T_2 di σ^2 è valido solo se l'ipotesi nulla H_0 è vera.

Il test confronta i valori di questi due stimatori e rifiuta l'ipotesi nulla H_0 quando il rapporto tra T_2 e T_1 è abbastanza grande.

12.2 Analisi della varianza ad un fattore

Consideriamo m campioni, ognuno di numerosità n , tra loro indipendenti, provenienti da popolazioni

normali.

Per $i = 1, \dots, m$, sia dato il campione i -esimo estratto dalla popolazione con media μ_i e varianza σ^2 .

Vogliamo verificare

l'ipotesi nulla $H_0 : \mu_1 = \mu_2 = \dots = \mu_m$

contro

l'ipot. altern. $H_1 : \text{non tutte le medie sono uguali}$

Per ogni campione i -esimo siano:

\bar{X}_i la sua media campionaria,

S_i^2 la sua varianza campionaria.

Ogni S_i^2 è uno stimatore corretto di σ^2 ($E[S_i^2] = \sigma^2$ per il teorema 2 del campionamento).

Poichè abbiamo m di questi stimatori, li combiniamo calcolandone la media:

$$T_1 = \frac{1}{m} \sum_{i=1}^m S_i^2$$

T_1 è uno stimatore corretto di σ^2 a prescindere dalla veridicità o meno di H_0 .

Supponiamo ora che H_0 sia vera, cioè:

$$\mu_i = \mu, \quad \forall i = 1, \dots, m$$

Ciò implica che le m medie campionarie $\bar{X}_1, \dots, \bar{X}_m$ sono tutte normali con la stessa media μ e la stessa varianza $\frac{\sigma^2}{n}$, cioè:

$$\bar{X}_i \sim N\left(\mu, \frac{\sigma^2}{n}\right), \quad \forall i = 1, \dots, m$$

Perciò la varianza campionaria,

$$\bar{S}^2 = \frac{1}{m-1} \sum_{i=1}^m (\bar{X}_i - \bar{\bar{X}})^2,$$

dove abbiamo posto $\bar{\bar{X}} = \frac{1}{m} \sum_{i=1}^m \bar{X}_i$, risulterà essere

uno stimatore corretto di $\frac{\sigma^2}{n}$, cioè:

$$E[\bar{S}^2] = \frac{\sigma^2}{n}.$$

Allora il secondo stimatore è:

$$T_2 = n\bar{S}^2$$

Riassumendo:

T_1 stima sempre σ^2

T_2 stima σ^2 solo se l'ipotesi nulla H_0 è vera.

La statistica test è data da:

$$ST = \frac{n\bar{S}^2}{\frac{1}{m} \sum_{i=1}^m S_i^2}$$

e H_0 è rifiutata quando ST è abbastanza grande.

Quando H_0 è vera ST ha come distribuzione una F di Fisher, ottenuta come rapporto di due χ^2 .
 numeratore = χ^2 con $(m - 1)$ gradi di libertà
 denominatore = χ^2 con $m(n - 1)$ gradi di libertà
 $F_{(m-1),m(n-1);\alpha}$ è il valore critico α di questa distribuzione, cioè:

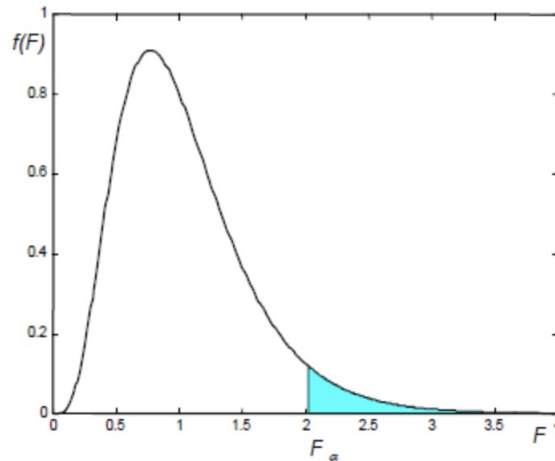


Figura 39: F di Fisher

$$P[F \geq F_{r,s;\alpha}] = \alpha.$$

Il test al livello α per H_0 è il seguente:

$$\text{rifiuta } H_0 \text{ se } \frac{n\bar{S}^2}{\frac{1}{m} \sum_{i=1}^m S_i^2} \geq F_{(m-1), m(n-1); \alpha}$$

accetta H_0 altrimenti.

12.3 Analisi della varianza a due fattori

Supponiamo che ogni valore dei dati sia influenzato da due fattori e che il primo fattore abbia m valori possibili o *livelli* e che il secondo fattore ne abbia n . Costruiamo una tabella di dati X_{ij} , $i = 1, \dots, m$, $j = 1, \dots, n$:

$$\begin{array}{ccc} X_{11} & \dots & X_{1n} \\ X_{21} & \dots & X_{2n} \\ \cdot & \dots & \cdot \\ \cdot & \dots & \cdot \\ X_{m1} & \dots & X_{mn} \end{array}$$

dove i e j sono rispettivamente il fattore riga e il fattore colonna.

X_{ij} sono v.c. normali, indipendenti con la stessa varianza σ^2 .

Il modello ANOVA a due fattori presuppone che

$$E[X_{ij}] = \mu + \alpha_i + \beta_j$$

dove

$\mu = \frac{1}{m} \sum_{i=1}^m \mu_i$ è la media generale

α_i = deviazione da μ dovuta alla riga i

β_j = deviazione da μ dovuta alla colonna j

con

$$\sum_{i=1}^m \alpha_i = \sum_{j=1}^n \beta_j = 0.$$

Gli stimatori dei parametri μ, α_i, β_j sono:

$$\hat{\mu} = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n X_{ij} = X_{rc} \text{ (media di tutti i valori } mn \text{ dei dati)}$$

$$\hat{\alpha} = \frac{1}{n} \sum_{j=1}^n X_{ij} - X_{rc}$$

$$\hat{\beta} = \frac{1}{m} \sum_{i=1}^m X_{ij} - X_{rc}$$

Uno stimatore corretto di σ^2 è dato da:

$$\frac{SS_e}{N}$$

dove $N = (n - 1)(m - 1)$ e

$$SS_e = \sum_{i=1}^m \sum_{j=1}^n (X_{ij} - X_{ic} - X_{rj} - X_{rc})^2$$

è la somma dei quadrati degli errori dove

$$X_{ic} = \frac{1}{n} \sum_{j=1}^n X_{ij}, \quad X_{rj} = \frac{1}{m} \sum_{i=1}^m X_{ij}.$$

Infine si definiscono:

$$SS_r = n \sum_{i=1}^m (X_{ic} - X_{rc})^2,$$

$$SS_c = n \sum_{j=1}^n (X_{rj} - X_{rc})^2$$

rispettivamente le somme dei quadrati delle righe e delle colonne.

Verifica delle ipotesi

1) Vogliamo verificare:

H_0 : tutte le α_i sono nulle

(il valore di un dato non è influenzato dal fattore riga)

contro

H_1 : non tutte le α_i sono nulle

La statistica del test è data dalla funzione:

$$ST = \frac{\frac{SS_r}{m-1}}{\frac{SS_e}{N}}$$

Il test al livello α è:

rifiuto H_0 se $ST > F_{m-1, N; \alpha}$
accetto H_0 altrimenti

2) Vogliamo verificare:

H_0 : tutti i β_j sono nulli

(il valore di un dato non è influenzato dal fattore colonna)

contro

H_1 : non tutti i β_j sono nulli

La statistica del test è data dalla funzione:

$$ST = \frac{\frac{SS_c}{n-1}}{\frac{SS_e}{N}}$$

Il test al livello α è:

rifiuto H_0 se $ST > F_{n-1, N; \alpha}$
accetto H_0 altrimenti