INTRODUZIONE ALL'ANALISI DISCRIMINANTE

Concetti preliminari

a) Scomposizione della matrice di Varianze e covarianze.

Si consideri una popolazione suddivisa in 2 gruppi.

n_1 => numero di unità statistiche estratte dal primo gruppo

n_2 => n° di unità statistiche estratte dal 2° gruppo

$n = n_1 + n_2$

Sulle n unità statistiche estratte vengono rilevati 2 caratteri: X e Y.

$x_{11}; x_{12}; \dots; x_{1n_1}$ => rilevazione di X sul gruppo 1

$x_{21}; x_{22}; \dots; x_{2n_2}$ => rilevazione di X sul gruppo 2

$y_{11}; y_{12}; \dots; y_{1n_1}$ => rilevazioni di Y sul gruppo 1

$y_{21}; y_{22}; \dots; y_{2n_2}$ => rilevazioni di Y sul gruppo 2.

Come è noto, la varianza di X (σ_x^2) è scomponibile in varianza NEI gruppi ed in varianza FRA i gruppi:

$$\sigma_x^2 = \underbrace{\frac{1}{n}\sum_{g=1}^{2} \sigma_{gx}^2 \cdot n_g}_{} + \underbrace{\frac{1}{n}\sum_{g=1}^{2} (\mu_{gx} - \mu_x)^2 n_g}_{}$$

Var "NEI" gruppi   Var "FRA" i gruppi

$$\sigma^2_{gx} = \frac{1}{n_g} \sum_{i=1}^{n_g} (x_{gi} - \mu_{gx})^2$$

$$\mu_{gx} = \frac{1}{n_g} \sum_{i=1}^{n_g} x_{gi} \qquad \mu_x = \frac{1}{n} \sum_{g=1}^{2} \mu_{gx} \cdot n_g$$

In maniera analoga si scompone la varianza di Y.

Si consideri ora la covarianza tra i caratteri X e Y.

$$Cov(X,Y) = \sigma_{xy} = \frac{1}{n} \sum_{g=1}^{2} \underbrace{\sum_{i=1}^{n_g} (x_{gi} - \mu_x)(y_{gi} - \mu_y)}$$

$$\sum_{i=1}^{n_g} (X_{gi} - \mu_x)(Y_{gi} - \mu_y) =$$

$$= \sum_{i=1}^{n_g} (X_{gi} - \mu_{gx} + \mu_{gx} - \mu_x)(y_{gi} - \mu_{gy} + \mu_{gy} - \mu_y)$$

$$= \sum_{i=1}^{n_g} \left[(X_{gi} - \mu_{gx}) + (\mu_{gx} - \mu_x) \right] \left[(y_{gi} - \mu_{gy}) + (\mu_{gy} - \mu_y) \right]$$

$$= \sum_{i=1}^{n_g} (x_{gi} - \mu_{gx})(y_{gi} - \mu_{gy}) + \underbrace{\sum_{i=1}^{n_g} (X_{gi} - \mu_{gx})(\mu_{gy} - \mu_y)}_{=0} +$$

$$+ \underbrace{\sum_{i=1}^{n_g} (y_{gi} - \mu_{gy})(\mu_{gx} - \mu_x)}_{=0} + \sum_{i=1}^{n_g} (\mu_{gx} - \mu_x)(\mu_{gy} - \mu_y)$$

$$= \sum_{i=1}^{n_g} (x_{gi} - \mu_{gx})(y_{gi} - \mu_{gy}) + n_g (\mu_{gx} - \mu_x)(\mu_{gy} - \mu_y)$$

Di conseguenza :

$$\sigma_{xy} = \frac{1}{n}\sum_{g=1}^{2}\left[\sum_{i=1}^{n_g}(x_{gi}-\mu_{gx})(y_{gi}-\mu_{gy}) + n_g(\mu_{gx}-\mu_x)(\mu_{gy}-\mu_y)\right]$$

$$= \frac{1}{n}\sum_{g=1}^{2}\sum_{i=1}^{n_g}(x_{gi}-\mu_{gx})(y_{gi}-\mu_{gy}) + \frac{1}{n}\sum_{g=1}^{2}(\mu_{gx}-\mu_x)(\mu_{gy}-\mu_y)\,n_g$$

$$= \underbrace{\frac{1}{n}\sum_{g=1}^{2}\sigma_{gxy}\,n_g}_{\text{covarianza nei gruppi}} + \underbrace{\frac{1}{n}\sum_{g=1}^{2}(\mu_{gx}-\mu_x)(\mu_{gy}-\mu_y)\,n_g}_{\text{covarianza fra i gruppi.}}$$

L'espressione della varianza tra i gruppi si può notevolmente semplificare, In dettaglio, si ricordi che:

$$\mu_x = \frac{\mu_{1x}\,n_1 + \mu_{2x}\,n_2}{n} \qquad \text{da cui}$$

- $$\mu_{1x}-\mu_x = \frac{1}{n}\left[\overbrace{(n-n_1)}^{n_2}\mu_{1x} - \mu_{2x}\,n_2\right]$$

$$= \frac{n_2}{n}(\mu_{1x}-\mu_{2x})$$

(risultati analoghi valgono per y)

- $$\mu_{2x}-\mu_x = \frac{n_1}{n}(\mu_{2x}-\mu_{1x})$$

Di conseguenza ⟶ var "FRA".

$$\frac{1}{n}\sum_{g=1}^{n}(\mu_{gx}-\mu_x)^2 n_g$$

$$= \frac{1}{n}\left[\frac{n_2^2}{n^2}(\mu_{1x}-\mu_{2x})^2 n_1 + \frac{n_1^2}{n^2}(\mu_{2x}-\mu_{1x})^2 n_2\right]$$

$$= \frac{n_1 n_2}{n^3}\left[n_2(\mu_{1x}-\mu_{2x})^2 + n_1(\mu_{1x}-\mu_{2x})^2\right]$$

$$= \frac{n_1 n_2}{n^2}(\mu_{1x}-\mu_{2x})^2$$

Allo stesso modo si ottiene:

$$\frac{1}{n}\sum_{g=1}^{2}(\mu_{gy}-\mu_y)^2 n_g = \frac{n_1 n_2}{n^2}(\mu_{1y}-\mu_{2y})^2$$

Relativamente alla covarianza fra i gruppi si ha:

$$\frac{1}{n}\left[(\mu_{1x}-\mu_x)(\mu_{1y}-\mu_y)n_1 + (\mu_{2x}-\mu_x)(\mu_{2y}-\mu_y)n_2\right] =$$

$$= \frac{n_1 n_2^2}{n^3}(\mu_{1x}-\mu_{2x})(\mu_{1y}-\mu_{2y}) +$$

$$+ \frac{n_1^2 n_2}{n^3}(\mu_{2x}-\mu_{1x})(\mu_{2y}-\mu_{1y}) =$$

$$= \frac{n_1 n_2}{n^2}(\mu_{1x}-\mu_{2x})(\mu_{1y}-\mu_{2y})$$

Ricapitolando:

$$\sigma_x^2 = \frac{1}{n}\left(\sigma_{1x}^2 n_1 + \sigma_{2x}^2 n_2\right) + \frac{n_1 n_2}{n^2}(\mu_{1x}-\mu_{2x})^2$$

$$\sigma_y^2 = \frac{1}{n}\left(\sigma_{1y}^2 n_1 + \sigma_{2y}^2 n_2\right) + \frac{n_1 n_2}{n^2}\left(\mu_{1y} - \mu_{2y}\right)^2$$

$$\sigma_{xy} = \frac{1}{n}\left(\sigma_{1xy} n_1 + \sigma_{2xy} n_2\right) + \frac{n_1 n_2}{n^2}\left(\mu_{1x} - \mu_{2x}\right)\left(\mu_{1y} - \mu_{2y}\right)$$

Matrice di Varianze - Covarianze di X ed Y.

$$S = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix} \Rightarrow$$ scomponibile additivamente

$$S = S_w + S_b$$

dove

$$S_w = \frac{n_1}{n}\begin{bmatrix} \sigma_{1x}^2 & \sigma_{1xy} \\ \sigma_{1xy} & \sigma_{1y}^2 \end{bmatrix} + \frac{n_2}{n}\begin{bmatrix} \sigma_{2x}^2 & \sigma_{2xy} \\ \sigma_{2xy} & \sigma_{2y}^2 \end{bmatrix}$$

$$S_b = \frac{n_1 n_2}{n^2}\begin{bmatrix} (\mu_{1x} - \mu_{2x})^2 & (\mu_{1x} - \mu_{2x})(\mu_{1y} - \mu_{2y}) \\ (\mu_{1x} - \mu_{2x})(\mu_{1y} - \mu_{2y}) & (\mu_{1y} - \mu_{2y})^2 \end{bmatrix}$$

$$= \frac{n_1 n_2}{n^2}\begin{bmatrix} \mu_{1x} - \mu_{2x} \\ \mu_{1y} - \mu_{2y} \end{bmatrix}\begin{bmatrix} \mu_{1x} - \mu_{2x} \\ \mu_{1y} - \mu_{2y} \end{bmatrix}^T$$

$$= \frac{n_1 n_2}{n^2} (\underline{\mu}_1 - \underline{\mu}_2)(\underline{\mu}_1 - \underline{\mu}_2)^T$$

dove $\underline{\mu}_i = \begin{bmatrix} \mu_{ix} \\ \mu_{iy} \end{bmatrix}$

È a questo punto banale generalizzare la scomposizione appena presentata al caso di q variabili.

$$\underset{q \times q}{S} = \underset{q \times q}{S_W} + \underbrace{\frac{n_1 n_2}{n^2} (\underline{\mu}_1 - \underline{\mu}_2)(\underline{\mu}_1 - \underline{\mu}_2)^T}_{q \times q}$$

dove

$$S_W = \frac{n_1}{n} S_1 + \frac{n_2}{n} S_2$$

$S_i \equiv$ matrice di varianze covarianze delle q variabili rilevate sull'i-esimo gruppo

$\underline{\mu}_i \equiv$ vettore delle medie delle q variabili rilevate sull'i-esimo gruppo.

osservazione: $\text{Rango}(S_b) = \underline{1}$

Concetti preliminari

β) Varianza della somma e Varianza di una combinazione lineare

Siano $X_1 \dots X_q$ dei caratteri quantitativi che vengono rilevati su n unità statistiche.
Si ottiene così la matrice di dati

$$\underset{n \times q}{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1q} \\ \vdots & & & & & \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{iq} \\ \vdots & & & & & \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{nq} \end{bmatrix}$$

dalla quale si deriva la matrice di varianze-covarianze

$$S = \frac{1}{n} X^T \left[I_n - \frac{1}{n} U_n \right] X$$

dove I_n indica la matrice identità di dimensione $n \times n$ mentre U_n indica la matrice $n \times n$ i cui elementi sono tutti pari ad 1.
Si è interessati ad ottenere la varianza del carattere

$$Y = \sum_{i=1}^{q} X_i$$

Come noto:

$$\sigma^2_{Y} = \sum_{i=1}^{q} \sigma^2_{X_i} + 2 \sum^{q} \sum^{q} \sigma_{X_i X_j}$$

$$\sigma_y^2 = \sum_{i=1}^{q} \sigma_{X_i}^2 + 2 \sum_{i=1} \sum_{j=i+1} \sigma_{X_i X_j}$$

$$= \mathbb{1}_q^T S \, \mathbb{1}_q$$

dove $\mathbb{1}_q$ è il vettore di q i cui elementi sono tutti pari ad 1.

Più in generale, se si è interessati alla varianza di

$$y^* = \sum_{i=1}^{q} a_i X_i ,$$

sfruttando le ben note proprietà della varianza e della covarianza, si ottiene:

$$\sigma_{y^*}^2 = \sum_{i=1}^{q} a_i^2 \sigma_{X_i}^2 + 2 \sum_{i=1}^{q} \sum_{j=i+1}^{q} a_i a_j \sigma_{X_i X_j}$$

$$= a^T S a$$

dove $a = \begin{bmatrix} a_1 \\ \vdots \\ a_i \\ \vdots \\ a_q \end{bmatrix}$.

OSSERVAZIONE: Utilizzando la notazione matriciale, i valori dei caratteri y e y^* corrispondenti alle rilevazioni sulle n unità statistiche si ottengono così:

$$\underset{n \times 1}{\underline{y}} = \underset{n \times q}{\underline{X}}\, \mathbb{1}_q$$

$$\underset{n \times 1}{\underline{y}^*} = \underset{n \times q}{\underline{X}}\, a$$

Concetti preliminari

8) Derivazione di una forma quadratica simmetrica.

$S \Rightarrow$ matrice $q \times q$ simmetrica di elementi s_{ij}

$$a^T S a = \sum_{i=1}^{q} \sum_{j=1}^{q} a_i a_j s_{ij}$$

$$= \sum_{i=1}^{q} a_i^2 s_{ii} + \sum_{i=1}^{q} \sum_{\substack{j=1 \\ i \neq j}}^{q} a_i a_j s_{ij}$$

$$= \sum_{i=1}^{q} a_i^2 s_{ii} + 2 \sum_{i=1}^{q} \sum_{j=i+1}^{q} a_i a_j s_{ij} \qquad (*)$$

Si è interessati a ricavare il vettore delle derivate parziali di $a^T S a$ rispetto agli elementi di a:

$$\frac{\partial a^T S a}{\partial a} = \begin{bmatrix} \dfrac{\partial a^T S a}{\partial a_1} \\ \vdots \\ \dfrac{\partial a^T S a}{\partial a_i} \\ \vdots \\ \dfrac{\partial a^T S a}{\partial a_q} \end{bmatrix}$$