Concetti chiave e regole

La dipendenza statistica

Nell'analisi di un fenomento statistico si cerca spesso di indagare sulla possibile dipendenza di due caratteri uno dall'altro.
La **teoria della correlazione** studia tale dipendenza nel caso in cui X e Y sono delle variabili statistiche, mettendo in rilievo, in particolare, se vi è dipendenza di tipo lineare.
L'indice che dà informazioni sulla dipendenza lineare è quello di **Bravais-Pearson** che è così definito:

$$\rho(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$$

essendo σ_X e σ_Y le deviazioni standard di X e di Y e $\text{cov}(X,Y) = \dfrac{\sum(x_i - \overline{x})(y_i - \overline{y})}{n}$
Se capita che:

- $\rho = \pm 1$ le due variabili sono perfettamente correlate e quindi, essendoci perfetta dipendenza lineare, i dati si distribuiscono su una retta
- $0 < \rho < 1$ le variabili sono correlate positivamente e più ρ si avvicina a 1, più la dipendenza lineare è forte, cioè i dati tendono a distribuirsi lungo una retta di coefficiente angolare positivo
- $-1 < \rho < 0$ le variabili sono correlate negativamente e più ρ si avvicina a -1, più i dati tendono a distribuirsi lungo una retta di coefficiente angolare negativo
- $\rho = 0$ le variabili non sono correlate, cioè non vi è dipendenza lineare, anche se non si possono escludere altri tipi di dipendenza.

L'interpolazione statistica

Una funzione di **interpolazione statistica** è una curva che passa fra i punti (x_i, y_i) corrispondenti ai dati rilevati.
Per scegliere la funzione che meglio approssima i dati, si applica il metodo dei minimi quadrati che consiste nello scegliere la funzione $f(x)$ che rende minima la quantità $S = \sum_{i=1}^{n}[f(x_i) - y_i]^2$ che esprime la somma dei quadrati delle distanze dei punti rilevati dalla funzione $f(x)$.

La funzione interpolante più semplice da determinare è la funzione lineare che in molte situazioni riesce ad approssimare sufficientemente bene i dati. Essa ha equazione $y = mx + q$ dove i coefficienti m e q sono espressi dalle formule

$$m = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \qquad q = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

Un metodo alternativo per determinare la retta di interpolazione statistica è quello del baricentro dove la retta ha equazione

$$y - \overline{y} = m(x - \overline{x}) \qquad \text{essendo:} \quad \overline{x} = \frac{\sum x_i}{n} \qquad \overline{y} = \frac{\sum y_i}{n} \qquad m = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}$$

Se i dati rilevati riguardano una serie storica, la retta di interpolazione può descrivere il comportamento tendenziale o **trend** della variabile Y per valori di X che vanno oltre quelli rilevati.

La retta di regressione

Il coefficiente di Bravais-Pearson indica se fra due variabili vi è la tendenza ad una dipendenza di tipo lineare, ma non dà modo di trovare l'espressione di questa dipendenza.
La teoria della **regressione** consente invece di esprimere la dipendenza della variabile X da Y, o viceversa, determinando rispettivamente la retta di regressione di X su Y o di Y su X; queste due rette si determinano con il metodo del baricentro oppure dei minimi quadrati ed hanno equazioni:

- regressione di Y su X: $y = ax + b$
- regressione di X su Y: $x = cy + d$

dove i coefficienti a, b, c, d si trovano con le formule relative all'interpolazione lineare, con l'avvertenza di scambiare i valori x_i con i valori y_i per la determinazione di c e d.