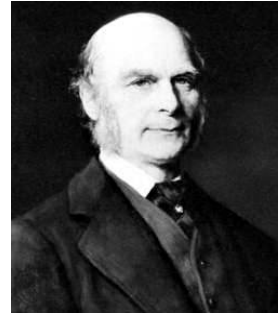


Un metodo di rappresentazione geometrica dei coefficienti di correlazione e di determinazione

Marco Piumetti

Il *coefficiente di correlazione* è un importante strumento statistico che consente di stabilire il grado di relazione lineare tra due serie di dati. Fu lo scienziato Francis Galton cugino di Charles Darwin, ad utilizzare per la prima volta il coefficiente di correlazione, proposto dal collega Karl Pearson, per quantificare il “grado di associazione” tra due variabili. Tale coefficiente, noto come coefficiente di Pearson, viene definito come il rapporto tra la sommatoria dei prodotti dei vari punteggi standard delle due variabili ed i gradi di libertà; esso può variare da un massimo di +1,00 (perfetta correlazione positiva) ad un minimo di -1,00 (perfetta correlazione negativa).



Francis Galton 1822-1911

$$r = \frac{1}{n-1} \sum \left(\frac{X_i - \bar{X}}{s_x} \right) \left(\frac{Y_i - \bar{Y}}{s_y} \right)$$

In cui: $\frac{X_i - \bar{X}}{s_x}$, \bar{X} , s_x

rappresentano rispettivamente i punteggi standard, la media e la deviazione standard (calcolata usando n-1 al denominatore).

Il quadrato del coefficiente di correlazione, meglio noto come *coefficiente di determinazione*, misura la “varianza spiegata”, ossia la percentuale della variabilità di Y spiegata dalla variabilità di X. Pertanto, tale coefficiente può essere molto utile per valutare l’influenza di una variabile su di un’altra. Se, per esempio, si considera il fatto che esiste una significativa correlazione tra la presenza di persone senza lavoro e la percentuale di disoccupati con un’educazione inferiore alla scuola secondaria ($r = 0,70$) ciò significa che circa il 50 % della varianza della disoccupazione potrebbe essere spiegato in termini di scolarità.

Il coefficiente di correlazione può essere rappresentato graficamente come il coseno dell’angolo formato tra due vettori associati alle variabili X ed Y. Calcolando, infatti, il prodotto scalare tra due vettori (X, Y) si può risalire all’angolo θ ed è pertanto possibile rappresentare nel piano due serie di dati.

$$X \cdot Y = \|Y\| \|X\| \cos \theta$$

Tuttavia, il coefficiente di correlazione che si ricava è generalmente non centrato (r^*), poiché la media delle due variabili può essere diversa da zero (condizione necessaria per ottenere r centrato).

$$\cos \theta = \frac{X \cdot Y}{\|X\| \|Y\|} = r^*$$

$$\theta = \arccos \left(\frac{X \cdot Y}{\|X\| \|Y\|} \right) = \arccos r^*$$

Per ottenere il coefficiente di correlazione centrato, cioè quello corretto, occorre calcolare il valor medio di ciascuna delle due serie X, Y e traslare tutti i dati rispetto a tali medie, affinché le due variabili X' , Y' siano entrambe centrate.

$$X' (X_1 - \bar{X} ; X_2 - \bar{X} ; \dots\dots\dots X_n - \bar{X})$$

$$Y' (Y_1 - \bar{Y} ; Y_2 - \bar{Y} ; \dots\dots\dots Y_n - \bar{Y})$$

Applicando, infine, la formula seguente si ricava l'angolo corrispondente al coefficiente di correlazione "corretto".

$$\theta = \arccos\left(\frac{X' \cdot Y'}{\|X'\| \|Y'\|}\right) = \arccos r$$

Esempio: X (1 2 4 5 7); Y (1 3 4 5 10) r: 0,953

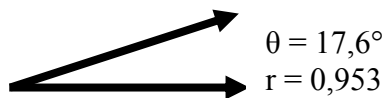
$$\cos \theta = \frac{118}{\sqrt{95}\sqrt{151}} = 0,985 = r^*$$

$$\bar{X} = 3,8 \quad \bar{Y} = 4,6$$

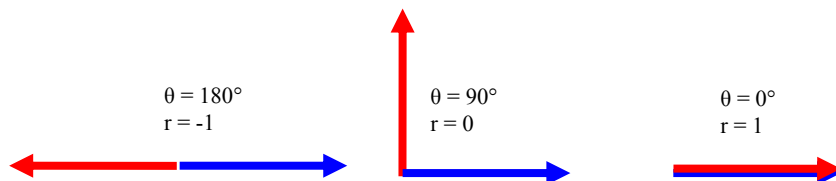
$$X' (-2,8 - 1,8 0,2 1,2 3,2) \quad Y' (-3,6 -1,6 -0,6 0,4 5,4)$$

$$\cos \theta = \frac{30,6}{\sqrt{22,8}\sqrt{45,2}} = 0,953 = r$$

$$\theta = \arccos r = \arccos 0,953 = 17,6^\circ$$



Nelle tre figure successive sono riportati, a titolo d'esempio, alcuni casi limite che si possono ottenere con la rappresentazione grafica del coefficiente di correlazione.



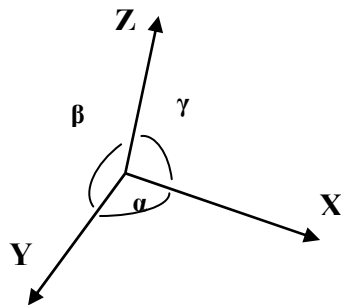
In modo del tutto analogo, si può rappresentare graficamente il coefficiente di determinazione per visualizzare meglio l'influenza di una variabile sull'altra, cioè la varianza spiegata.

Anche in tal caso, attraverso dei semplici calcoli trigonometrici, si può ricavare la relazione tra l'angolo θ formato tra i due vettori associati alle due variabili ed il coefficiente di determinazione.

$$\theta = \arccos r^2$$

Per esempio, ipotizzando una significativa correlazione ($r = 0,8$) tra i voti ottenuti dagli studenti all'esame di biologia e le ore di studio, si ottiene che circa il 64 % dei voti può essere spiegato in termini di tempo dedicato alla preparazione dell'esame.

Quando si hanno tre variabili correlate tra di loro è possibile rappresentare geometricamente i coefficienti di correlazione e determinazione mediante dei vettori disposti nello spazio secondo gli angoli α, β, γ calcolati per ciascuna coppia di variabili.



Inoltre, quando le tre variabili X, Y, Z risultano tutte correlate ad una variabile principale P è possibile rappresentarle geometricamente mediante dei vettori in modo tale che il vettore associato alla variabile principale P sia disposto secondo gli angoli α, β, γ formati con i tre vettori corrispondenti rispettivamente alle variabili X, Y, Z.

