Il coefficiente di correlazione lineare r di Bravais-Pearson

Riassunto

Il coefficiente di correlazione lineare r di Bravais-Pearson è costruito come media geometrica dei coefficienti di regressione esposti nella rubrica "Statistica per concetti" del precedente numero di CASCO. Si mostra come esso possa essere utile per risolvere il problema del confronto della concordanza tra due o più distribuzioni doppie. Infine, si accenna agli indici di concordanza tra due graduatorie da usare in luogo di r quando almeno uno dei caratteri associati sia qualitativo rettilineo (cioè ordinale): il rho di Spearman e il tau di Kendall.

Parole chiave. *Coefficienti di regressione, coefficiente di correlazione lineare, r, ρ (rho) di Spearman, τ (tau) di Kendall.*

Summary

Pearson's correlation coefficient

Pearson's correlation coefficient, r, is calculated as geometric mean of the two regression coefficients described in the previous number of CASCO, and it is shown that r can be used to compare the concordance (or discordance) between two or more double distributions. Furthermore, when at least one characteristic is evaluated using an ordinal scale, r should be replaced by Spearman's ϱ or by Kendall's τ.

Key words. *Regression coefficient, Pearson's correlation coefficient, r, Spearman's ρ, Kendall's τ.*

Nel precedente numero di CASCO a partire dalla rappresentazione grafica della nuvola di punti sono state definite le due rette di regressione come quelle che passano il più vicino possibile ai punti empirici (X, Y) in base al metodo dei minimi quadrati. Più precisamente, la prima, quella di Y a X, ha equazione Y = a + bX i cui parametri sono calcolati rendendo minima la somma dei quadrati delle distanze tra ordinate teoriche (quelle dei punti che giacciono sulla retta) e ordinate empiriche (quelle della nuvola di punti). La seconda, quella di X a Y, ha equazione X = a' + b'Y, ed i suoi parametri sono ottenuti minimizzando la somma dei quadrati degli scarti tra ascisse teoriche (quelle dei punti appartenenti alla retta) e ascisse empiriche (quelle dei punti della nuvola): v. grafici nel precedente numero di CASCO.

Si è mostrato che i coefficienti angolari, b e b', sono misure di concordanza che hanno lo stesso segno e variano da −∞ a +∞. Se il segno dei coefficienti angolari è positivo si dice che c'è concordanza, nel senso che al crescere di X (ad es. la statura) in media cresce anche Y (ad es., il peso: al crescere del peso cresce, in media, la statura e al crescere della statura cresce, in media il peso; in altre parole i soggetti più alti sono anche quelli che, in media, sono i più pesanti). Analogamente, se il segno dei coefficienti angolari è negativo, c'è discordanza, ossia al crescere di un carattere, l'altro, in media, diminuisce.

Si è infine mostrato che il valore del coefficiente angolare di entrambe le rette di regressione è un indice di concordanza, nel senso che misura ***quanto*** *cresce un carattere, in media, al crescere di una unità dell'altro carattere (ad es., se in un gruppo di ragazzi di 10 anni, l'equazione della retta di regressione della statura rispetto al peso fosse X = 40 + 1,2Y, il coefficiente angolare (+1,2) indicherebbe che la statura cresce, in media, di 1,2 cm per ogni aumento di un chilo di peso).*

In conclusione, sono stati costruiti due indici di concordanza, uno di Y a X, l'altro di X a Y.

Nel caso si voglia confrontare la concordanza tra due o più distribuzioni doppie, avere due indici, b e b', non è conveniente perché il risultato del confronto potrebbe dipendere dalla scelta dell'uno o dell'altro; inoltre essi variano da −∞ a +∞ mentre in Statistica, per eseguire confronti, si preferiscono gli indici "normalizzati" ossia quelli che assumono valori in un range ben definito (ad es., tra 0 e 1).

Si possono sintetizzare b e b' in un solo valore facendone una media. Quella più vantaggiosa è la media geometrica che, in questo caso, è la radice quadrata del prodotto tra b e b'. Essa prende il nome di coefficiente di correlazione lineare di Bravais-Pearson (dal nome degli autori che, indipendentemente l'uno dall'altro, lo hanno introdotto) e il suo simbolo è r:

$$r = \pm \sqrt{(b \times b')}.$$

I due coefficienti di regressione hanno lo stesso segno; pertanto il loro prodotto è sempre positivo. Com'è noto, il risultato dell'operazione di radice quadrata può essere assunto sia con segno positivo che con segno negativo (ad es., $\sqrt{4} = 2$ ed anche $\sqrt{4} = -2$ perché l'operazione inversa, il quadrato, in ambedue i casi vale 4).

Poiché r è la media geometrica di b e b', ed ogni media è interna ai termini della distribuzione, r dovrà essere compreso tra b e b'. Il segno di radice è preceduto dal simbolo "±" perché si sceglie il segno "+" se i coefficienti di regressione sono entrambi positivi e il segno "−" se sono negativi (altrimenti la media risulterebbe esterna).

Calcolato come media geometrica di b e b', r ha il notevole vantaggio di variare tra – 1 e + 1 (di essere normalizzato) denotando con i valori negativi discordanza e concordanza con quelli positivi. Nel caso in cui è r = 0, sono anche b = 0 e b' = 0; tale caso è detto di "indifferenza": al variare di X, Y non varia lungo la retta di regressione di Y a X e, al variare di Y, X resta costante sulla retta di regressione di X a Y.

Ad esempio, in un gruppo di ragazzi maschi di terza media, la retta di regressione della statura (Y) rispetto al peso (X) ha equazione Y = 20 + 1,4X e quella del peso rispetto alla statura X = –10 + 0,4Y. Pertanto, è r = $\sqrt{(1,4 \times 0,4)} = \sqrt{0,56} = 0,748$. Interpretazione: in quel gruppo di ragazzi, al crescere di 1 chilo di peso, la statura cresce in media (cioè lungo la retta di regressione) di 1,4 cm; al crescere di 1 cm di statura, il peso cresce in media di 0,4 kg. Il valore di r = 0,748 indica che in questa distribuzione doppia c'è il 74,8% della massima concordanza che vi si sarebbe potuta osservare. È evidente il vantaggio di r per i confronti; ad esempio se volessimo confrontare la concordanza tra peso e statura dei maschi (r = 0,748) con quella riscontrata nel collettivo delle loro colleghe (ad es., r = 0,589) si concluderebbe che al crescere del peso la statura cresce in modo più marcato per gli studenti maschi che per le loro colleghe, così come al crescere della statura, il peso cresce maggiormente per i maschi che per la femmine.

Quanto esposto è applicabile quando i due caratteri associati (ad es., peso e statura) sono entrambi quantitativi. Spesso però, in Medicina, un carattere è quantitativo, mentre l'altro è qualitativo ordinale (ad es., in soggetti neoplastici, score di qualità di vita e tempo dalla diagnosi di malattia); ovvero sono entrambe scale ordinali (ad es., due sottoscale di un questionario psicometrico per la misura della qualità di vita). In tali casi, per misurare la concordanza occorre ricorrere ad indici non parametrici perché i coefficienti di regressione non possono essere calcolati. Il più usato indice non parametrico corrispondente ad r è il ϱ (rho) di Spearman (esistono anche altri indici non parametrici di concordanza, come il τ (tau) di Kendall, ma si incontrano più raramente). Tutti tali indici, calcolati da tutti i *package* statistici, sono detti "indici di concordanza tra graduatorie" ed hanno le stesse proprietà e la stessa interpretazione di r: variano tra – 1 e + 1, denotando con valori negativi discordanza, con valori positivi concordanza e con zero indifferenza.

Enzo Ballatori