Statistica inferenziale
Test di ipotesi

Bari, 7 Gennaio 2008

Coefficiente di correlazione

Abbiamo visto come misurare l'associazione tra variabili aleatorie discrete. L'analisi della correlazione mira a misurare l'associazione tra v.a. di tipo continuo, posto che tra esse vi sia una relazione di tipo lineare.

Ricordiamo che il coefficiente di correlazione ρ tra le v.a. X e Y é definito come:

$$\rho_{xy} = \frac{\mathbb{E}[(X - \mu_x)(Y - \mu_y)]}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

dove $\mu_x = \mathbb{E}(X)$ e $\mu_y = \mathbb{E}(Y)$. Inoltre $-1 \leq \rho_{xy} \leq 1$ e se $\rho_{xy} = 0$ allora le v.a. X e Y sono non correlate.

Coefficiente di correlazione

Consideriamo un campione accoppiato costituito da n osservazioni delle v.a. $X : N(\mu_x, \sigma_x)$ e $Y : N(\mu_y, \sigma_y)$:

$$(X_1, Y_1), ..., (X_n, Y_n).$$

Uno stimatore di ρ é il coefficiente di correlazione di Pearson dato da:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

Ipotesi nulla e alternativa

Vogliamo testare le seguenti ipotesi:

$$\begin{aligned} H_0 &: (\rho = 0)\ X \text{ e } Y \text{ sono non correlate;} \\ H_1 &: (\rho \neq 0)\ X \text{ e } Y \text{ sono correlate.} \end{aligned}$$

Si puó dimostrare che, sotto H_0 vera, la variabile:

$$T_0 = r\sqrt{\frac{n-2}{1-r^2}}$$

é una t di Student con $(n-2)$ gradi di libertá $t(n-2)$.

Regione critica e test

La regione critica D di livello α é data da:

$$D = \left\{ |T_0| > t_{1-\frac{\alpha}{2}}(n-2) \right\}.$$

Esecuzione del test:

- calcola il valore empirico t_0 di T_0;
- confronta t_0 con il quantile $t_{1-\frac{\alpha}{2}}(n-2)$;
- se $|t_0| > t_{1-\frac{\alpha}{2}}(n-2)$ allora rigetta H_0 e accetta H_1, altrimenti accetta H_0.

Esempio

Esaminare la relazione tra la percentuale di bambini vaccinati all'etá di un anno contro DPT in un Paese ed il corrispondente tasso di mortalitá in bambini al di sotto dei 5 anni in quel Paese.

Paese	% Vaccinati	Morti ogni 1000 nati vivi	Paese	% Vaccinati	Morti ogni 1000 nati vivi
Bolivia	77	118	*Giappone*	87	6
Brasile	69	65	*Grecia*	54	9
Cambogia	32	184	*India*	89	124
Canada	85	8	*Italia*	95	10
Cina	94	43	*Messico*	91	33
Egitto	89	55	*Polonia*	98	16
Etiopia	13	208	*RegnoUnito*	90	9
Fed.Russa	73	32	*Rep.Ceca*	99	12
Finlandia	95	7	*Senegal*	47	145
Francia	95	9	*Turchia*	76	87

Esempio (cont.)

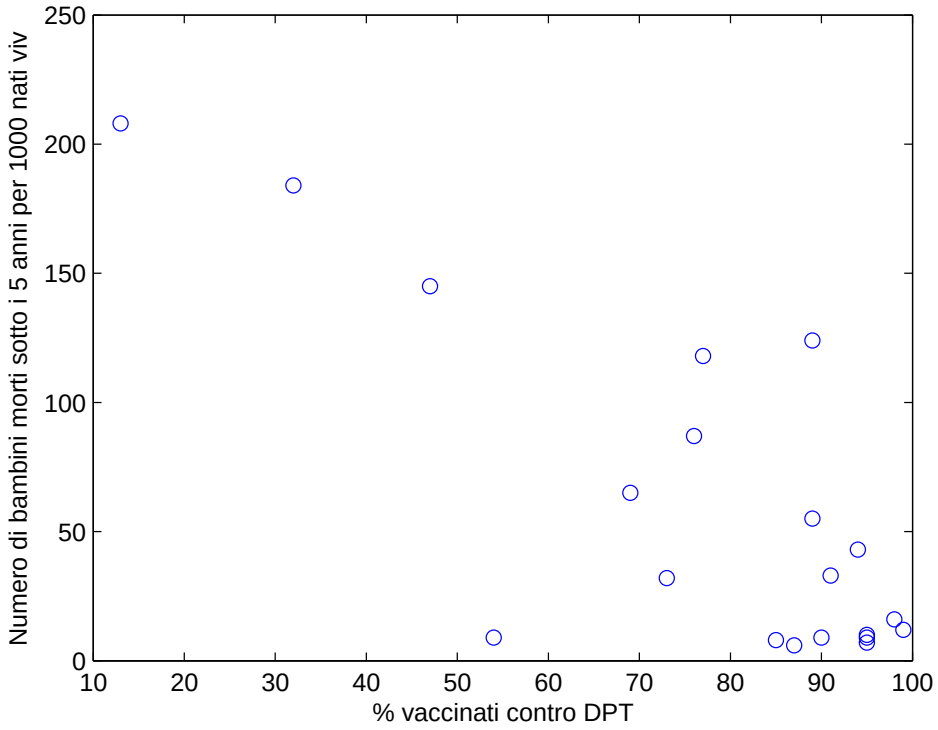

Figure 1: Diagramma a punti dei dati in tabella.

Esempio (cont.)

Calcoliamo le medie $\overline{x} = 77.4$, $\overline{y} = 59.0$ e le varianze $s_x^2 = 559.5$, $s_y^2 = 4078.8$. Inoltre $\sum_{i=1}^{20}(x_i - 77.4)(y_i - 59.0) = -22706$. Pertanto:

$$r = \frac{-22706}{\sqrt{(n-1)s_x^2}\sqrt{(n-1)s_y^2}} = \frac{-22706}{(19)(23.6541)(63.8658)} = -0.7911.$$

Esempio (cont.)

Per $\alpha = 0.05$, confrontiamo il valore di

$$t_0 = r\sqrt{\frac{n-2}{1-r^2}} = -0.7911\sqrt{\frac{20-2}{1-(-0.7911)^2}} = -5.4870$$

con il quantile $t_{0.975}(20-2) = 2.1009$. Poiché $|t_0| > 2.1009$ allora a livello α rigettiamo l'ipotesi nulla e concludiamo che esiste una forte correlazione tra la percentuale di bambini vaccinati ed il tasso di mortalitá al di sotto di 5 anni.

Regressione

Vogliamo affrontare il problema di predire o stimare il valore di una v.a. Y a partire dalla conoscenza del valore x assunto da una v.a. X.

Per esempio vogliamo predire il peso P di una persona conoscendo la sua altezza h.

Oppure, vogliamo analizzare il tipo di relazione esistente tra due v.a.

In generale le v.a. X e Y sono continue. Quando Y é discreta allora si parla di classificazione.

Funzione di regressione

Supponiamo di conoscere la funzione densitá di probabilitá congiunta $p(x, y)$ delle due v.a. X e Y. Allora si definisce "funzione di regressione" f la funzione:

$$f(x) = \mathbb{E}(Y|x) = \int_{-\infty}^{+\infty} yp(y|x)dy.$$

Nel caso in cui le v.a. X e Y sono Normali congiuntamente, allora la funzione di regressione é una retta.

Regressione lineare

Consideriamo un campione accoppiato $(x_1, y_1), ..., (x_n, y_n)$. Determiniamo i coefficienti w e b della retta $y = wx + b$ che meglio approssima i dati. A tale scopo consideriamo l'errore quadratico medio:

$$\epsilon(w, b) = \frac{1}{n} \sum_{i=1}^{n} (y_i - (wx_i + b))^2$$

Determiniamo i coefficienti w^* e b^* che minimizzano l'errore quadratico medio. Questa fase viene detta "training".

Calcolo dei coefficienti

A tale scopo poniamo uguale a zero le derivate parziali di ϵ rispetto a w e b:

$$\frac{\partial \epsilon}{\partial w} = 0 \Rightarrow$$

$$-\frac{2}{n}\sum_{i=1}^{n}(y_i - (wx_i + b))x_i = 0 \Rightarrow \sum_{i=1}^{n}(y_i x_i - wx_i^2 - bx_i) = 0 \Rightarrow$$

$$\sum_{i=1}^{n} y_i x_i = w\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i. \tag{1}$$

Calcolo dei coefficienti (cont.)

Analogamente:

$$\frac{\partial \epsilon}{\partial b} = 0 \Rightarrow$$

$$-\frac{2}{n}\sum_{i=1}^{n}(y_i - (wx_i + b)) = 0 \Rightarrow \sum_{i=1}^{n}(y_i - wx_i - b) = 0 \Rightarrow$$

$$\sum_{i=1}^{n} y_i = w\sum_{i=1}^{n} x_i + nb. \qquad (2)$$

Le equazioni (1) e (2) costituiscono un sistema lineare di 2 equazioni nelle incognite w e b.

Calcolo dei coefficienti (cont.)

Da (2) segue che $b = \overline{y} - w\overline{x}$. Sostituendo in (1) si ha:

$$\sum_{i=1}^{n} y_i x_i = w \sum_{i=1}^{n} x_i^2 + (\overline{y} - w\overline{x}) \sum_{i=1}^{n} x_i \Rightarrow$$

$$\sum_{i=1}^{n} y_i x_i = w \sum_{i=1}^{n} x_i^2 + \overline{y} \sum_{i=1}^{n} x_i - w\overline{x} \sum_{i=1}^{n} x_i \Rightarrow$$

$$w \left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2 \right) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i$$

Calcolo dei coefficienti (cont.)

Da cui:

$$w^* = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \tag{3}$$

Sostituendo w^* in $b = \overline{y} - w\overline{x}$ si ha:

$$b^* = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}. \tag{4}$$

Errore di training

Una volta calcolati w^* e b^* che minimizzano l'errore quadratico medio, possiamo calcolare l'errore di training:

$$\epsilon(w^*, b^*) = \frac{1}{n} \sum_{i=1}^{n} (y_i - (w^* x_i + b^*))^2$$

Piú piccolo é ϵ, meglio i dati sono descritti da una retta.

Errore di predizione

Supponiamo di voler conoscere il valore y associato ad un "nuovo" x, ossia ad un dato non presente nel nostro campione. Stimiamo y attraverso $\widehat{y} = w^* x + b^*$. Qual é l'errore medio che si commette quando stimiamo y con $\widehat{y}$?

Si definisce errore di predizione o errore di generalizzazione la quantitá:

$$\mathbb{E}[(y - \widehat{y})^2] = \int (y - (w^* x + b^*))^2 \, p(x, y) dx dy$$

Poiché la densitá $p(x, y)$ é incognita, allora l'errore di predizione non si puó calcolare.

Leave-One-Out Error (Luntz and Brailovsky 1969)

Una stima dell'errore di generalizzazione che non fa uso della conoscenza della densitá di probabilitá dei dati é basata sulla seguente procedura: per ogni $i = 1, 2, ..., n$:

- escludi l'esempio (x_i, y_i) dal campione;
- calcola w_i^* e b_i^* sui dati rimanenti $(x_1, y_1), ..., (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), ..., (x_n, y_n)$;
- calcola l'errore $\epsilon_i = (y_i - (w_i^* x_i + b_i^*))^2$.

Allora $LOOE = \frac{1}{n} \sum_{i=1}^{n} \epsilon_i$ é una stima non distorta dell'errore di generalizzazione.