Statistica inferenziale
Test di ipotesi

Bari, 17 Dicembre 2007

Test del χ^2 di adattamento

Affrontiamo ora il problema di decidere se un campione di n osservazioni puó essere considerato estratto da una v.a. X distribuita secondo una determinata legge. Ossia:

dato un campione casuale $X_1, X_2, ..., X_n$ di n osservazioni, possiamo affermare che il campione é stato estratto da una v.a. X con una ben precisa distribuzione di probabilitá?

Ipotesi nulla ed alternativa

Supponiamo che la v.a. X assuma un numero finito m di valori con probabilitá teoriche $p_1, p_2, ..., p_m$. Le ipotesi del test sono:

$$\begin{array}{lcl} H_0 & : & \text{il campione segue la legge di probabilitá } p_1, p_2, ..., p_m; \\ H_1 & : & \text{il campione segue un'altra legge.} \end{array}$$

Poiché nel test i valori assunti dalla X non hanno alcuna importanza, il test puó essere applicato anche a v.a. qualitative. Inoltre, il test puó essere applicato a v.a. che assumono infiniti valori o a v.a. continue. A tale scopo é sufficiente raggruppare i valori assunti dalla v.a. in un numero finito di classi.

Descrizione del test

Indichiamo con N_j il numero di volte che nel campione osservato la v.a. X assume il j-mo valore, con $j = 1, 2, ..., m$. Inoltre, siano $\overline{p}_j = N_j/n$ le corrispondenti frequenze relative. Nota che N_j e $\overline{p}_j$ sono v.a. mentre p_j sono numeri. Il test consiste nel confrontare le p_j teoriche con le $\overline{p}_j$ empiriche. A tale scopo consideriamo la statistica:

$$D_0 = n \sum_{j=1}^{m} \frac{(\overline{p}_j - p_j)^2}{p_j} = \sum_{j=1}^{m} \frac{(N_j - np_j)^2}{np_j}.$$

Regione critica del test

Si puó dimostrare che, se H_0 é vera e se n é grande, allora D_0 segue una legge $\chi^2(m-1)$ con $m-1$ gradi di libertá.

Se H_0 é vera la statistica D_0 assume valori piccoli, mentre se H_1 é vera le differenze $\overline{p}_j - p_j$ non sono piú trascurabili e D_0 assume valori grandi. La regione critica di un test di livello α é:

$$D = \{D_0 > \chi^2_{1-\alpha}(m-1)\}$$

dove $\chi^2_{1-\alpha}(m-1)$ é il quantile di ordine $1-\alpha$ della legge χ^2 con $m-1$ gradi di libertá.

Esecuzione del test

1. calcola il valore empirico d_0 di D_0;

2. confronta d_0 con il quantile $\chi^2_{1-\alpha}(m-1)$, dove m é il numero di possibili valori di X;

3. se $d_0 > \chi^2_{1-\alpha}(m-1)$ allora rigetta H_0 e accetta H_1, altrimenti accetta H_0.

L'uso del test richiede alcune precisazioni.

Precisazione 1

Deve sempre risultare $np_j \geq 5$ per ogni $j = 1, 2, ..., m$, dove n é la dimensione del campione. Se questa proprietá non é verificata, il procedimento piú comune consiste nel raggruppare le classi con frequenze N_j piú piccole in modo da formare classi piú grandi che soddisfano la proprietá $np_j \geq 5$.

Precisazione 2

In alcuni casi la distribuzione teorica p_j non é completamente nota. Ad esempio vogliamo stabilire se i nostri dati si adattano ad una legge $B(n, p)$ dove p non é nota. In questo caso, prima si stimano i parametri necessari a partire dai dati $X_1, X_2, ..., X_n$ e dopo si applica una versione del test cosí modificata:

detto q il numero di parametri da stimare per specificare completamente la distribuzione teorica, bisogna prima stimare questi parametri dai dati e poi si applica il test considerando la regione critica:

$$D = \{D_0 > \chi^2_{1-\alpha}(m - q - 1)\}$$

Esempio

Consideriamo un campione costituito da $n = 6115$ osservazione di una v.a. X che assume solo i valori $j = 0, 1, ..., 12$.

j	N_j	$\overline{p}_j$
0	3	0.00049
1	24	0.00392
2	104	0.01701
3	286	0.04677
4	670	0.10957
5	1033	0.16893
6	1343	0.21962
7	1112	0.18185
8	829	0.13557
9	478	0.07817
10	181	0.02960
11	45	0.00736
12	7	0.00114

Vogliamo stabilire se X segue una legge Binomiale $B(12, \frac{1}{2})$.

Esempio (cont.)

Esplicitiamo le probabilitá p_j teoriche nel caso $X : B(12, \frac{1}{2})$.

j	N_j	$\overline{p}_j$	p_j
0	3	0.00049	0.00024
1	24	0.00392	0.00293
2	104	0.01701	0.01611
3	286	0.04677	0.05371
4	670	0.10957	0.12085
5	1033	0.16893	0.19336
6	1343	0.21962	0.22559
7	1112	0.18185	0.19336
8	829	0.13557	0.12085
9	478	0.07817	0.05371
10	181	0.02960	0.01611
11	45	0.00736	0.00293
12	7	0.00114	0.00024

dove $p_j = \left(\begin{array}{c} 12 \\ j \end{array} \right) \left(\frac{1}{2}\right)^j \left(1 - \frac{1}{2}\right)^{12-j}$.

Esempio (cont.)

Osserva che $np_0 = np_{12} = 6115 \times 2^{-12} = 1.49 < 5$ e quindi la condizione 1 non é verificata. A tale scopo unifichiamo la classe $\{X = 0\}$ con la classe $\{X = 1\}$, e la classe $\{X = 12\}$ con la classe $\{X = 11\}$. In questo modo abbiamo: $n(p_0 + p_1) = 19.47 > 5$ e $n(p_{11} + p_{12}) = 19.47 > 5$.

Dobbiamo ora eseguire un test del χ^2 non sulla Binomiale originaria, ma su quella raggruppata nelle 11 classi con probabilitá:

$$q_j = \begin{cases} p_0 + p_1 & \text{per } j = 1, \\ p_j & \text{per } j = 2, ..., 10, \\ p_{11} + p_{12} & \text{per } j = 11. \end{cases}$$

Esempio (cont.)

Avendo raggruppato i dati in classi le frequenze assolute per le classi raggruppate diventano:

$$M_j = \begin{cases} N_0 + N_1 & \text{per } j = 1, \\ N_j & \text{per } j = 2, ..., 10, \\ N_{11} + N_{12} & \text{per } j = 11. \end{cases}$$

e con $m = 11$ si calcola la statistica: $d_0 = \sum_{j=1}^{m} \frac{(M_j - nq_j)^2}{nq_j} = 242.05$.

Confrontando d_0 con il quantile $\chi^2_{1-\alpha}(m-1) = \chi^2_{0.95}(10) = 18.31$, allora rigettiamo H_0 e concludiamo che la distribuzione empirica $\overline{p}_j$ non si adatta ad una $B(12, \frac{1}{2})$.

Esempio (cont.)

Verifichiamo ora l'ipotesi che la distribuzione di X sia $B(12, \overline{p})$ dove $\overline{p}$ é una stima di p ottenuta dai dati sperimentali.

Se $X : B(12, \overline{p})$ allora $\mathbb{E}(X) = 12\overline{p}$. Se stimiamo $\mathbb{E}(X)$ con la media campionaria $\overline{X}$, allora:

$$\overline{p} = \frac{\overline{X}}{12} = 0.51922.$$

Ora bisogna ripetere la stessa procedura tenendo in mente che il valore di p non é $\frac{1}{2}$, ma 0.51922.

Esempio (cont.)

Esplicitiamo le probabilitá p_j teoriche nel caso X : $B(12, 0.51922)$.

j	N_j	$\overline{p}_j$	p_j
0	3	0.00049	0.0002
1	24	0.00392	0.0020
2	104	0.01701	0.0117
3	286	0.04677	0.0423
4	670	0.10957	0.1027
5	1033	0.16893	0.1775
6	1343	0.21962	0.2236
7	1112	0.18185	0.2070
8	829	0.13557	0.1397
9	478	0.07817	0.0671
10	181	0.02960	0.0217
11	45	0.00736	0.0043
12	7	0.00114	0.0004

dove $p_j = \left(\begin{array}{c} 12 \\ j \end{array} \right) (0.51922)^j \, (1 - 0.51922)^{12-j}$.

Esempio (cont.)

Raggruppando i dati in classi si giunge ad un valore della statistica $d_0 = 105.79$. Confrontando questo valore con il quantile $\chi^2_{1-\alpha}(m-1-1) = \chi^2_{0.95}(9) = 16.918$, allora rigettiamo H_0 e concludiamo che la distribuzione empirica $\overline{p}_j$ non si adatta ad una distribuzione binomiale.

Nota che é stato sottratto un grado di libertá in quanto abbiamo dovuto stimare dai dati un parametro per poter definire la distribuzione teorica.

Test del χ^2 di indipendenza

Consideriamo due v.a. X e Y e supponiamo che esse possono assumere un numero finito di valori (anche di tipo qualitativo) $u_1, ..., u_r$ e $v_1, ..., v_s$ rispettivamente. Vogliamo stabilire se le due v.a. sono indipendenti.

A tale scopo misuriamo le due v.a. su n soggetti distinti, ottenendo un campione aleatorio accoppiato

$$(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n).$$

Tabella di contingenza

Indichiamo con N_{jk} il numero di volte in cui nel campione compare la coppia di valori (u_j, v_k). Inoltre indichiamo con $N_{j.}$ e con $N_{.k}$ il numero di volte in cui compare il valore u_j e v_k rispettivamente. Riportiamo i dati in tabella:

		Y				
		v_1	v_2	...	v_s	Tot
	u_1	N_{11}	N_{12}	...	N_{1s}	$N_{1.}$
	u_2	N_{21}	N_{22}	...	N_{2s}	$N_{2.}$
X	⋮	⋮	⋮		⋮	⋮
	u_r	N_{r1}	N_{r2}	...	N_{rs}	$N_{r.}$
	Tot	$N_{.1}$	$N_{.2}$	...	$N_{.s}$	n

Calcolo delle frequenze empiriche

Stimiamo i valori delle probabilitá congiunte

$$p_{jk} = \mathbb{P}\{(X, Y) = (u_j, v_k)\}$$

e delle probabilitá marginali

$$p_j = \mathbb{P}\{X = u_j\} \qquad q_k = \mathbb{P}\{Y = v_k\}$$

dai dati disponibili utilizzando le frequenze relative:

$$\overline{p}_{jk} = \frac{N_{jk}}{n} \qquad \overline{p}_j = \frac{N_{j.}}{n} \qquad \overline{q}_k = \frac{N_{.k}}{n} \qquad j = 1, ..., r \;\; k = 1, ..., s.$$

Confronto delle frequenze empiriche

Se X e Y sono indipendenti allora la probabilitá congiunta é uguale al prodotto delle probabilitá marginali: $p_{jk} = p_j q_k$. Per cui ha senso considerare un test che confronta le frequenze congiunte empiriche $\overline{p}_{jk}$ con i prodotti $\overline{p}_j \overline{q}_k$. A tale scopo consideriamo la statistica:

$$D_0 = n \sum_{j,k} \frac{(\overline{p}_{jk} - \overline{p}_j \overline{q}_k)^2}{\overline{p}_j \overline{q}_k} = \sum_{j,k} \frac{(N_{jk} - n\overline{p}_j \overline{q}_k)^2}{n\overline{p}_j \overline{q}_k}.$$

Osserva che dalle formule delle frequenze si ha:

$$n\overline{p}_j \overline{q}_k = \frac{N_{j.} N_{.k}}{n}$$

Test

Si dimostra che se le frequenze relative marginali non sono troppo piccole, nel caso di indipendenza delle v.a. (H_0), la statistica D_0 segue una legge $\chi^2[(r-1)(s-1)]$ con $(r-1)(s-1)$ gradi di libertá.

La regione critica di livello α é:

$$\{D_0 > \chi^2_{1-\alpha}[(r-1)(s-1)]\}.$$

Il test si puó applicare anche a v.a. continue raggruppando in classi i loro possibili valori.

Esempio

Relazione tra infezione HPV ed stadi di HIV: sintomativo (PS) e asintomatico (PA).

		HIV			
		PS	PA	N	Tot
HPV	P	23	4	10	37
	N	10	14	35	59
	Tot	33	18	45	96

Esempio (cont.)

Costruiamo la tabella nel caso di indipendenza tra le due infezioni mantenendo fisse le quantitá marginali utilizzando la relazione $n\overline{p}_j\overline{q}_k = \frac{N_{j.}N_{.k}}{n}$:

		HIV			
		PS	PA	N	Tot
HPV	P	$23(12.72)$	$4(6.94)$	$10(17.34)$	37
	N	$10(20.28)$	$14(11.06)$	$35(27.66)$	59
	Tot	33	18	45	96

Esempio (cont.)

Calcoliamo la statistica D_0:

$$D_0 = \frac{(23-12.72)^2}{12.72} + \frac{(4-6.94)^2}{6.94} + ... + \frac{(35-27.66)^2}{27.66} = 20.60081.$$

Confrontiamo D_0 con il quantile di ordine $1-\alpha$ di una distribuzione χ^2 con $(r-1)(s-1) = (2-1)(3-1) = 2$ gradi di libertá. Per $\alpha = 0.05$ si ha: $\chi^2_{0.95}(2) = 5.99146$. Poiché $D_0 > \chi^2_{0.95}(2)$ allora a livello $\alpha = 0.05$ rigettiamo H_0 e quindi le due v.a. sono dipendenti, ossia esiste una relazione tra HPV e gli stadi di HIV.

Tabella di contingenza 2×2

Spesso si ha a che fare con v.a. dicotomiche, ossia che assumono due soli valori. In questo caso avremo tabelle di contingenza 2×2:

	Y		Tot
X	a	b	$a+b$
	c	d	$c+d$
Tot	$a+c$	$b+d$	n

Tabella di contingenza 2×2 (cont.)

Calcoliamo:

$$D_0 = \frac{\left(a - \frac{(a+b)(a+c)}{n}\right)^2}{\frac{(a+b)(a+c)}{n}} + \frac{\left(b - \frac{(a+b)(b+d)}{n}\right)^2}{\frac{(a+b)(b+d)}{n}} + \frac{\left(c - \frac{(c+d)(a+c)}{n}\right)^2}{\frac{(c+d)(a+c)}{n}} + \frac{\left(d - \frac{(c+d)(b+d)}{n}\right)^2}{\frac{(c+d)(b+d)}{n}}$$

Semplificando si ha:

$$D_0 = \frac{n(ad - bc)^2}{(a+c)(b+d)(a+b)(c+d)}.$$

In questo caso, D_0 sotto H_0 vera segue una legge $\chi^2(1)$ con 1 grado di libertá.

Tabella di contingenza 2×2: correzione di Yates

Per problemi di discretizzazione, nelle tabelle 2×2 si preferisce correggere il valore di D_0 nel seguente modo:

$$D_0' = \frac{n(|ad - bc| - 0.5n)^2}{(a+c)(b+d)(a+b)(c+d)}.$$

Tale correzione, riducendo il valore di D_0, rende il test molto conservativo.

Esempio

Su 129 soggetti, 103 sono stati trattati con antibiotico. Un batterio é risultato essere multiresistente in 37 dei 129 soggetti. Vogliamo stabilire se c'é una relazione tra la condizione di multiresistenza e l'assunzione di antibiotico.

		Isolato batterio		
		Si	No	Tot
A	Si	36	67	103
	No	1	25	26
	Tot	37	92	129

Esempio (cont.)

Calcoliamo:

$$D_0' = \frac{129(|(36)(25) - (67)(1)| - 0.5(129))^2}{(37)(92)(103)(26)} = 8.3575.$$

Sotto H_0, D_0' segue una legge $\chi^2(1)$ con 1 grado di libertá. A livello $\alpha = 0.05$, confrontiamo questo valore con $\chi^2_{0.95}(1) = 3.841$. Poiché $D_0 > 3.841$ allora dobbiamo rigettare H_0 e concludere che esiste una relazione tra le due variabili.

Intervallo di Confidenza per Odds Ratio

Ricordiamo che l'Odds Ratio (OR) é il rapporto tra l'odds (pronostico) in favore della malattia tra i soggetti esposti e l'odds in favore della malattia tra i soggetti non esposti:

$$OR = \frac{\frac{\mathbb{P}\{D|\text{esposto}\}}{1-\mathbb{P}\{D|\text{esposto}\}}}{\frac{\mathbb{P}\{D|\text{non esposto}\}}{1-\mathbb{P}\{D|\text{non esposto}\}}}$$

Anche in questo caso abbiamo due criteri di classificazione dei nostri dati: sano, malato e esposto, non esposto.

Tabella di contingenza 2×2

Riportiamo i nostri dati in una tabella 2×2:

	Esposti	Non esposti	Tot
D	a	b	$a+b$
H	c	d	$c+d$
Tot	$a+c$	$b+d$	n

e otteniamo le seguenti stime:

$$\mathbb{P}\{D|\text{esposto}\} = \frac{a}{a+c} \qquad 1 - \mathbb{P}\{D|\text{esposto}\} = \frac{c}{a+c}$$

$$\mathbb{P}\{D|\text{non esposto}\} = \frac{b}{b+d} \qquad 1 - \mathbb{P}\{D|\text{non esposto}\} = \frac{d}{b+d}$$

Stimatore di OR

Otteniamo il seguente stimatore dell'OR:

$$\widehat{OR} = \frac{[a/(a+c)]/[c/(a+c)]}{[b/(b+d)]/[d/(b+d)]} = \frac{a/c}{b/d} = \frac{ad}{bc}.$$

Si puó dimostrare che il $\log \widehat{OR}$ ha una distribuzione approssimativamente normale.

Intervallo di confidenza al 95%

L'intervallo di confidenza al 95% di $\log \widehat{OR}$ é:

$$CI_{\log \widehat{OR}}95\% = [\log \widehat{OR} - 1.96\sigma, \log \widehat{OR} + 1.96\sigma]$$

dove:

$$\sigma = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Mentre il CI per l'odds ratio é:

$$CI95\% = [e^{\log \widehat{OR} - 1.96\sigma}, e^{\log \widehat{OR} + 1.96\sigma}]$$

Esempio

Si vuole stabilire se il monitoraggio elettronico fetale durante il parto influenzi la frequenza di parti cesarei.

	Esposizione		
Parto cesareo	Si	No	Tot
Si	358	229	587
No	2492	2745	5237
Tot	2850	2974	5824

Esempio (cont.)

L'odds in favore del parto cesareo tra i soggetti esposti a monitoraggio rispetto al gruppo non sottoposto a monitoraggio é stimato come:

$$\widehat{OR} = \frac{ad}{bc} = \frac{(358)(2745)}{(229)(2492)} = 1.72 \Rightarrow \log \widehat{OR} = 0.542.$$

$$\sigma = \sqrt{\frac{1}{358} + \frac{1}{229} + \frac{1}{2492} + \frac{1}{2745}} = 0.089.$$

$$CI_{\log \widehat{OR}} 95\% = [\log \widehat{OR} - 1.96\sigma, \log \widehat{OR} + 1.96\sigma]$$

$$CI_{\log \widehat{OR}} 95\% = [0.542 - 1.96 * 0.089, 0.542 + 1.96 * 0.089] = [0.368, 0.716].$$

$$CI 95\% = [e^{0.368}, e^{0.716}] = [1.44, 2.05].$$

Nota che 1 é esterno all'intervallo. Quindi al 95% il pronostico di parto cesareo tra i feti sottoposti a monitoraggio é 1.44-2.05 maggiore del pronostico di parto cesareo tra i feti non sottoposti a monitoraggio.