Ratio Mathematica 11
(1996), 63 - 73

ATTUALI SVILUPPI NELLO STUDIO DELL'ASSOCIAZIONE TRA VARIABILI DI TIPO QUALITATIVO

Carlo Antonelli*, Enzo Ballone**

SUNTO - L'analisi statistica di variabili di tipo qualitativo fa riferimento alla funzione chi-quadrato. I primi sviluppi di questa funzione sono dovuti a Pearson (1904) e a Fisher (1929); successivamente è stata arricchita di validi contributi apportati da Yates, da Mantel-Haenszel ecc.
In questa nota si illustreranno alcune implementazioni della funzione chi-quadrato, in particolare come test statistico per la verifica dell'ipotesi di indipendenza e come misura dell'associazione lineare tra una variabile di tipo ordinale ed una di tipo dicotomico nell'ambito della sua scomposizione.

ABSTRACT - The most common procedure for analyzing contingency table data is by using chi-square statistic. The early development of chi-square analysis of contingency table is credited to Pearson (1904) and Fisher (1929), successively expanded by Yates, Mantel-Haenszel etc. In this paper some developments of the chi-square function has been outlined, particularly as statistical test for the null hypothesis of independence, for subdividing contingency tables and the chi-square test for linear association between ordinal variables.

* I.T.C. "R. de Sterlich" - Chieti;
** Cattedra di Statistica Medica, Univ. Degli Studi "G. D'Annunzio" - Chieti

1. INTRODUZIONE

Nell'analisi statistica di innumerevoli situazioni pratiche attinenti fenomeni del campo socio-sanitario, e non solo, ci si imbatte sempre più spesso nello studio di variabili di tipo qualitativo, ossia caratteri le cui modalità sono espresse da "attributi", "etichette" ecc che non sempre presentano un ordinamento oggettivo come ad esempio il sesso, la diagnosi, il tipo di intervento ecc. A questa tipologia di caratteri, detti variabili categoriali, appartengono anche quelle variabili quantitative continue che, talvolta per convenienza, sono considerate discrete, cioè costituite da un numero limitato di modalità. Ad esempio la variabile peso può essere raggruppata per classi a ciascuna delle quali viene assegnato un numero intero positivo.

In tali situazioni la metodologia statistica più usata per lo studio delle associazioni tra variabili fa riferimento alla funzione χ^2 di Pizzetti-Pearson; tale metodologia permette una valutazione in termini probabilistici dell'associazione tra i caratteri osservati.

Il presente lavoro, dopo aver richiamato i principali aspetti metodologici della distribuzione χ^2, riporta alcune "derivazioni" della funzione, e precisamente quelle che le attribuiscono una valenza di test statistico per la verifica di ipotesi, in particolare "la scomposizione del chi-quadrato" e "la verifica dell'ipotesi di linearità". Queste due implementazioni permettono, quando se ne presentano le condizioni, di approfondire lo studio della relazione individuata, isolando le modalità che maggiormente o esclusivamente realizzano l'associazione (scomposizione del chi-quadrato) oppure qualificando la tipologia della relazione stessa (verifica dell'ipotesi di linearità).

2. ALCUNE CONSIDERAZIONI METODOLOGICHE

Si consideri lo spazio dei campioni di numerosità n conseguente ad una estrazione casuale da una popolazione P, di cui si conosce la struttura rispetto a due variabili statistiche (per struttura si intende la conoscenza delle proporzioni di unità classificate in base a tutte le possibili coppie di modalità dei due caratteri osservati). Si verifica che molti campioni presentano una struttura uguale o "vicina" a quella della popolazione e pochi una struttura molto diversa.

Una misura di quanto ciascun campione si discosta dalla struttura della popolazione di provenienza è data dalla statistica CHI-QUADRATO nella sua espressione:

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(n_{ij} - np_{ij})^2}{np_{ij}} \qquad (1)$$

in cui:

- c ed r indicano il numero delle modalità delle due variabili;
- n_{ij} la frequenza assoluta, registrata nel campione, relativa alla coppia ij di modalità delle due variabili;
- n la numerosità del campione;
- p_{ij} la proporzione (frequenza relativa) di unità statistiche che presentano, nella popolazione, la coppia di modalità ij;
- np_{ij} la frequenza assoluta teorica del campione, relativa alla coppia ij di modalità, sotto l'ipotesi che esso presenti la stessa struttura della popolazione.

In riferimento all'espressione (1) si può osservare che essa non dipende dal tipo (qualitativo o quantitativo) di carattere in esame e, inoltre, si verifica che $\sum_{i=1}^{c}\sum_{j=1}^{r} n_{ij} = \sum_{i=1}^{c}\sum_{j=1}^{r} np_{ij} = n$. Quest'ultima eguaglianza implica che i k (k=cxr) scarti $l_{ij} = n_{ij} - np_{ij}$ non sono fra di loro indipendenti ma ognuno di essi è vincolato dalla condizione che sommato agli altri $k-1$ deve rendere nulla la somma $\sum_{i=1}^{c}\sum_{j=1}^{r} l_{ij}$. Questa condizione rappresenta il *vincolo* (v). Il numero delle modalità della variabile statistica diminuito del numero di vincoli costituisce il *grado di libertà (g)*, dove $g = k-v$.

La statistica chi-quadrato, che in sostanza misura la discrepanza fra le frequenze osservate nel campione e quelle "teoriche", risulterà uguale a zero ($\chi^2 = 0$) solo se $n_{ij} \equiv np_{ij}$ (caso di perfetta aderenza tra struttura del campione e struttura della popolazione), maggiore di zero ($\chi^2 > 0$) quanto più le frequenze osservate si discostano da quelle attese.

La distribuzione campionaria della statistica chi-quadrato è espressa dalla funzione di densità della variabile casuale del III tipo di Pearson che presenta la seguente espressione:

$$P(\chi^2 / g) = \frac{1}{2^{\frac{g}{2}}\Gamma(\frac{g}{2})} e^{-\frac{1}{2}\chi^2} (\chi^2)^{\frac{g}{2}-1} d(\chi^2) \quad \text{con } g=k-v \qquad (2)$$

essendo la funzione gamma data da: $\Gamma(x) = \int_0^{+\infty} t^{(x-1)} e^{-t} dt$. (3)

Graficamente la forma della curva si modifica al variare dei gradi di libertà, presenta un andamento esponenziale negativo per $g<3$ e successivamente una forma campanulare con asimmetria destra che, molto marcata per valori bassi di g, tende a ridursi all'aumentare dei gradi di libertà (per $g>30$ la distribuzione si approssima a quella gaussiana).

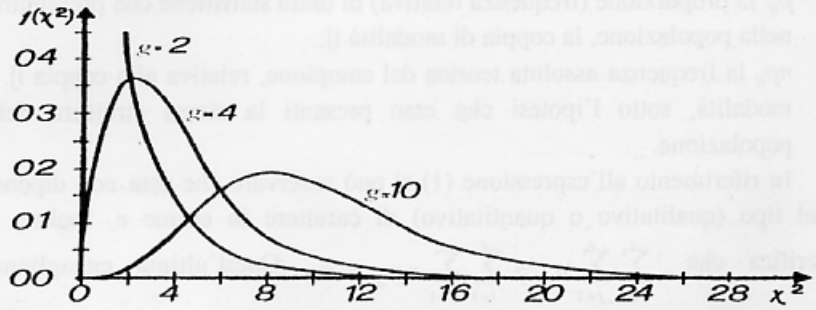

Fig.1 Funzione di distribuzione del χ^2 per alcuni gradi di libertà

3. IL χ^2 COME TEST STATISTICO PER LA VERIFICA DELL'IPOTESI DI INDIPENDENZA

Si ipotizzi che, in una popolazione P, due variabili siano indipendenti. Sotto questa ipotesi, chiamata usualmente ipotesi nulla (H_0), si è in grado di determinare le frequenze teoriche np_{ij} del campione e quindi, tramite la (1), il valore del χ^2 campionario (χ^2_o). Mediante la (2) si è in grado di calcolare la probabilità di ottenere un valore di χ^2 che, soddisfacendo ai ν vincoli imposti, sia uguale o maggiore a quello ottenuto dai dati del campione, ossia:

$$P(\chi^2 \geq \chi_o^2) = \frac{1}{2^{\frac{g}{2}}\Gamma(\frac{g}{2})} \cdot \int_{\chi_o^2}^{\infty} e^{-\frac{1}{2}\chi^2} (\chi^2)^{\frac{g}{2}-1} d(\chi^2) \quad (4)$$

Si può ovviare al calcolo dell'integrale, essendo disponibili delle tavole che forniscono, per prefissati gradi di libertà, valori teorici di χ^2 che possono essere superati con diversi livelli di probabilità, di solito si considerano il 5%, l'1% e l'1‰, denotando che solamente il 5% o l'1% o l'1‰ dei campioni, estratti casualmente dalla popolazione P in cui le due variabili sono indipendenti, presenterà un valore del χ^2 maggiore o uguale al valore "critico" tabulato (χ^2_c). In definitiva se dai dati del campione risulta un valore del $\chi^2_o \leq \chi^2_c$ il test non è significativo e lo scostamento tra ipotesi fatta e dati campionari è da attribuire a fattori casuali e/o di campionamento e si è indotti a non rigettare H_0; se, invece, $\chi^2_o > \chi^2_c$ il test risulta statisticamente significativo e segnala che si è verificata una delle due seguenti possibili alternative:

- il campione, ad esempio con prob. 0.05 (livello di significatività), è uno fra quel 5% di campioni non rappresentativi provenienti da una popolazione in cui è vera H_0;
- il campione, con prob. 0.95 (livello di fiducia), proviene da una popolazione in cui H_0 non è vera.

Naturalmente in tale situazione si accredita la seconda possibilità e si dirà che, con un livello di fiducia del 95%, si rigetta l'ipotesi che le variabili in esame siano indipendenti; ciò equivale ad ammettere, con un margine di errore del 5%, l'esistenza di un legame associativo fra i caratteri osservati. Il margine d'errore, all'occorrenza, può essere ridotto aumentando di conseguenza il grado di fiducia.

4. SCOMPOSIZIONE DEL CHI-QUADRATO

Una volta rigettata l'ipotesi di indipendenza H_0 in molti casi può essere utile approfondire l'analisi cercando di individuare le modalità delle due variabili per cui si manifesta la massima associazione. Ciò equivale ad analizzare i "contributi" che le singole modalità forniscono al legame associativo complessivo.

E' questo il problema della "scomposizione del χ^2" che viene risolto sfruttando la proprietà additiva della funzione.

Qualsiasi tabella con *r* righe e *c* colonne (*rxc*), con *r* e/o *c* >2, può essere ripartita in tante sottotabelle quanti sono i gradi di libertà della tabella di partenza ($g=(r-1)(c-1)$); ad esempio per tabelle *r*x2 si possono formare *r-1* sottotabelle 2x2 con il seguente criterio:

- la prima sottotabella si ottiene considerando solamente le prime due righe;
- le successive si ricavano dalla precedente "accorpando" le due righe presenti ed aggiungendo la riga successiva della tabella di partenza.

I totali marginali, per tener conto del fatto che queste tabelle 2x2 sono sottotabelle ottenute da una più grande, sono le distribuzioni marginali dell'intero campione. Naturalmente la ripartizione della tabella iniziale nelle sottotabelle 2x2 deve avere valenza logica, ossia deve rispondere ad un preordinato piano di confronti singoli programmato dal ricercatore. Di seguito si riporta uno schema di scomposizione per una tabella 3x2:

X \ Y	y_1	y_2	tot.
X_1	n_{11}	n_{12}	R_1
x_2	n_{21}	n_{22}	R_2
x_3	n_{31}	n_{32}	R_3
tot.	C_1	C_2	N

X \ Y	y_1	y_2	tot.
x_1	n_{11}	n_{12}	R_1
x_2	n_{21}	n_{22}	R_2
tot.	C_1	C_2	N

(a)

X \ Y	y_1	y_2	tot.
x_1+x_2	$n_{11}+n_{21}$	$n_{12}+n_{22}$	R_1+R_2
x_3	n_{31}	n_{32}	R_3
tot.	C_1	C_2	N

(b) (5)

Anche il test χ^2 per l'indipendenza, da utilizzare per ciascuna sottotabella, deve essere modificato per tener conto del fatto che si opera su sottotabelle ottenute da tabelle più grandi e quindi devono riflettere le caratteristiche del campione completo. A.W.Kimball (1954) ha proposto una serie di formule relative a vari gradi di libertà. Di seguito si riportano le formule del test χ^2 relative alle due sottotabelle estratte da una tabella 3x2, la formula generale per tabelle rx2 e la formula generale per generiche tabelle rxc:

$$\chi^2_{(a)} = \frac{N^2(n_{22}n_{11} - n_{21}n_{12})^2}{C_1C_2R_2R_1(R_1 + R_2)} \tag{6}$$

$$\chi^2_{(b)} = \frac{N[n_{32}(n_{11} + n_{21}) - n_{31}(n_{12} + n_{22})]^2}{C_1C_2R_3(R_1 + R_2)} \tag{7}$$

$$\chi^2_{(rx2)} = \frac{N^2(n_{r+1,2}\sum_{i=1}^{r} n_{i1} - n_{r+1,1}\sum_{i=1}^{r} n_{i2})^2}{C_1C_2R_{r+1}(\sum_{i=1}^{r} R_i)(\sum_{i=1}^{r+1} R_i)} \tag{8}$$

$$\chi^2_{(rxc)} = \frac{N[R_1(n_{21}C_2 - n_{22}C_1) - R_2(n_{11}C_2 - n_{12}C_1)]^2}{R_1R_2(R_1 + R_2)C_1C_2(C_1 + C_2)} \tag{9}$$

Esempio - Si ipotizzi che la distribuzione di 319 alunni del biennio di un istituto superiore, classificati in base alla distanza fra la loro residenza e la scuola (misurata in tempi di percorrenza), ed ai loro esiti scolastici in termini di successo (promozione) e insuccesso (bocciatura o abbandono), sia quella riportata nella tabella seguente:

	Esito scolastico		
Tempo (min.)	successo	insuccesso	Tot.
0-30	150	21	171
30-60	78	17	95
>60	37	16	53
Totale	265	54	319

Ci si può chiedere se, in base ai dati registrati, il "pendolarismo" influenzi negativamente gli esiti scolastici; cioè se le differenze di successi riscontrate tra i ragazzi pendolari e quelli residenti sono dovute al "caso" oppure dalla condizione di "pendolarismo".

Il test statistico χ^2 , secondo la formula seguente:

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(n_{ij} - n_{ij}^*)^2}{n_{ij}^*} \qquad (10)$$

dove n_{ij} sono le frequenze assolute osservate e n_{ij}^* le corrispondenti frequenze assolute "attese" nella ipotesi di " indipendenza " fra le variabili (H_0: l'esito scolastico è indipendente dai tempi di percorrenza), permette di dare una risposta al problema. La distribuzione delle frequenze attese, sotto H_0, è stata ottenuta applicando le percentuali marginali di riga al totale di ogni colonna della tabella precedente.

	Esito scolastico		
Tempo (min.)	successo	insuccesso	Tot.
0-30	142	29	171
30-60	79	16	95
>60	44	9	53
Totale	265	54	319

Utilizzando la (10) si ottiene $\chi^2_o = 9.32$ che, essendo maggiore di 7.82 (valore critico del χ^2 con 2 g ad un livello di fiducia del 95%), permette di affermare, con un margine di errore del 5%, l'esistenza di una associazione tra tempi di percorrenza ed esiti scolastici e quindi di concludere che "il pendolarismo" influisce negativamente sull'esito scolastico.

Volendo approfondire l'analisi per indagare su come i vari tempi di percorrenza vanno ad incidere sugli esiti scolastici, si ripartisce la tabella, in base allo schema (5), in due sottotabelle 2x2 che riportano gli esiti scolastici relativi ai tempi di percorrenza 0-30 / 30-60 e 0-60 / >60:

	Esito		
Tempo	suc.	ins.	
0-30	150	21	171
30-60	78	17	95
	265	54	319

	Esito		
Tempo	suc.	ins.	
0-60	228	38	266
>60	37	16	53
	265	54	319

Applicando la (6) e (7) alle due sottotabelle si ha:

$$\chi^2_{(a)} = \frac{319^2(17\cdot 150 - 78\cdot 21)^2}{265\cdot 54\cdot 95\cdot 171(171+95)} = 1.369$$

$$\chi^2_{(b)} = \frac{319[16(150+78) - 37(21+17)]^2}{265\cdot 54\cdot 53(171+95)} = 7.948$$

Il risultato del primo test risulta chiaramente non significativo, mentre quello del secondo evidenzia una significatività statistica del 5%; si può concludere, quindi, che solamente una distanza superiore ai 60 minuti di percorrenza influenza in maniera statisticamente significativa l'andamento scolastico degli alunni. La correttezza della procedura di scomposizione è confermata dalla relazione $\chi^2_{Tot} = \chi^2_{(a)} + \chi^2_{(b)}$.

5. SCOMPOSIZIONE DEL "CHI-QUADRATO PER IL TREND"

Il χ^2 come test statistico di indipendenza non fornisce informazioni sul tipo di relazione e/o associazione esistente tra le variabili in esame. Qualora la variabile risposta sia di tipo dicotomico (vero, falso; deceduto, sopravvissuto; migliorato, non migliorato ecc) e la variabile indipendente sia di tipo ordinale è possibile, attraverso la logica della scomposizione, suddividere il valore complessivo del test in due componenti; quella dovuta all'associazione di tipo lineare tra i due caratteri ed quella dovuta ai cosiddetti "fattori casuali". Naturalmente affinché ci sia una relazione di tipo lineare tra i due caratteri dovrà risultare significativa la componente lineare e non significativa quella dovuta al "caso". L'eventuale significatività di quest'ultima denoterebbe che ci sono altre variabili che influenzano il fenomeno, variabili che non vengono spiegate dal carattere considerato indipendente. Un esempio potrà chiarire l'importanza di questo tipo di analisi.

Esempio - Al fine di cogliere ulteriori elementi predittivi sul rendimento scolastico degli studenti, ai 319 ragazzi esaminati precedentemente è stato chiesto il numero medio di ore dedicate giornalmente allo studio. I risultati sono riportati nella seguente tabella:

	Ore di studio					
Esito scolastico	≤ 1	2	3	4	> 4	Tot.
Successo	10	43	74	104	34	265
Insuccesso	11	26	14	3		54
Totale	21	69	88	107	34	319

Scomposizione del Chi-quadrato totale:

Componenti del χ^2	Valore	g	p
χ^2_{Tot}	62.1	4	< 0.001
χ^2_{Lin}	57.7	1	< 0.001
χ^2_{Res}	4.4	3	n.s.

dove p indica la probabilità di ottenere valori di χ^2, sotto l'ipotesi di indipendenza, maggiori o uguali di quello campionario.

L'analisi delle componenti del test evidenzia come l'associazione tra le due variabili sia di tipo lineare, cioè al crescere delle ore di studio aumenta la probabilità per lo studente di essere promosso. Si può concludere, quindi, che il tempo dedicato giornalmente allo studio risulta essere un fattore "determinante" il rendimento scolastico.

6. CONSIDERAZIONI CONCLUSIVE

La funzione χ^2 di Pizzetti-Pearson assume, in definitiva, un ruolo di primo piano nella Statistica applicata allorquando, nell'ambito di una ricerca campionaria, si affronta lo studio delle associazioni intercorrenti tra due variabili categoriali. Pertanto si è voluto focalizzare l'attenzione su tale funzione-test richiamando i principali aspetti metodologici della sua distribuzione e riportando alcune derivazioni del test, quali la "scomposizione del chi-quadrato" e la "verifica dell'ipotesi di linearità", al fine di ampliare il campo di utilizzo di tale funzione. Queste implementazioni consentono una maggiore qualificazione della eventuale relazione individuata tra le variabili e permettono, quindi, uno studio più dettagliato sul "dove" e "come" l'associazione si realizza.

BIBLIOGRAFIA

1. A.W. KIMBALL, *Short-cut Formulas for the Exact Partitioning of chi-square in Contingency Table*, Biometrics 16, 1954.
2. P. ARMITAGE, *Statistica medica*, ed. Feltrinelli, Milano, 1992.
3. E. BALLATORI, *Sui test statistici per il confronto tra due frequenze in tabelle 2x2*, Metron, vol XL, 1982.
4. A. RIDOMI, *Una formula generale per la scomposizione del* χ^2, Statistica applicata, vol.4 - num.3, 1992.
5. G. CHISCI, *Biometria principi e metodi*, ed. Piccin, Padova, 1977.