Il test di chi-quadro*

Germano Rossi†

18 novembre 2004
vers. 0.6

Indice

*Questa dispensa è un capitolo di un lavoro più vasto intitolato *Elementi di ragionamento statistico: per psicologia e scienze dell'educazione* da cui è tratto.

†Università degli Studi di Milano-Bicocca, Dipartimento di Psicologia (germano.rossi@unimib.it)

1 Il test di chi-quadro [0.6]

1.1 Introduzione

Il test di chi-quadro (χ^2) è una *tecnica* di inferenza statistica che si basa sulla *statistica* di chi-quadro e sulla relativa distribuzione di probabilità[1]. Si può usare con variabili a livello di scala nominale e/o ordinale, generalmente disposte in forma di tabelle di contingenza.

Lo scopo principale di questa statistica è di verificare le differenze tra valori osservati e valori teorici (generalmente chiamati "attesi") e di effettuare un'inferenza sul grado di scostamento fra i due. Praticamente la tecnica viene usata con 3 diversi obiettivi, tutti basati sullo stesso principio fondamentale:

a) la casualità della distribuzione di una variabile categoriale;
b) l'indipendenza di due variabili qualitative (nominali o ordinali);
c) le differenze con un modello teorico.

Per lo scopo del punto b), l'indice statistico di chi-quadro può essere considerato come una statistica di associazione.

1.1.1 Casualità della distribuzione di una variabile

Per ora, ci limiteremo a considerare il primo aspetto (ovvero il punto *a*), utilizzando una ipotesi di lavoro della psichiatria, di qualche anno fa.

Per diversi anni, gli psichiatri hanno avanzato l'ipotesi che i pazienti affetti da schizofrenia nascano prevalentemente nei periodi invernali (Bradbury & Miller, 1985). Vogliamo vedere se anche i dati in nostro possesso ci portano a conclusioni analoghe. A questo scopo, usando le cartelle cliniche (del tutto inventate), raccogliamo le informazioni relative alla data di nascita di un certo numero di pazienti schizofrenici e li suddividiamo in categorie corrispondenti alle 4 stagioni dell'anno:

Tabella 1: Numero di schizofrenici nati nelle singole stagioni (Dati fittizi)

	Primavera	Estate	Autunno	Inverno	Totale
Soggetti schizofrenici	125	130	153	153	636

Se formulassimo un'ipotesi di assoluta uniformità, cioè che la stagione di nascita non ha nulla a che fare con il fatto di manifestare successivamente un patologia schizofrenica, dovremmo aspettarci che ogni cella contenga più o meno la stessa percentuale di nati (poiché ci sono 4 celle, il 25% circa, cioè 159). In pratica quello che faremmo è utilizzare un'ipotesi nulla di equiprobabilità:

$$H_0 : P(P) = P(E) = P(A) = P(I) = 0.25$$

contro un'ipotesi alternativa non equiprobabile

$$H_1 : P(P) \neq P(E) \neq P(A) \neq P(I) \neq 0.25$$

e di usare l'ipotesi nulla per generare le frequenze teoriche.

[1] Il termine "chi-quadro" si usa per indicare contemporaneamente la distribuzione di probabilità, una particolare tecnica di inferenza statistica e un indice statistico. Alcuni autori usano il simbolo χ^2 (chi greca minuscola) per indicare la distribuzione di probabilità e X^2 (chi greca maiuscola) per indicare la statistica. In questa sede, però, non faremo questa distinzione.

E' ben difficile però ottenere in ogni cella esattamente il valore atteso e si otterranno invece valori leggermente diversi che oscilleranno attorno a quello considerato uniforme. Valori molto vicini a quelli teorici avranno buone probabilità di essere delle "variazioni casuali", mentre valori molto diversi e lontani da quelli teorici avranno poche probabilità di essere considerati "fluttuazioni casuali". Serve quindi un criterio per decidere fino a che punto dobbiamo accettare come casuali le varie oscillazioni.

Il procedimento di calcolo che adotteremo è abbastanza simile a quello che abbiamo usato per la varianza e può essere riassunto concretamente così:

1.	Calcoliamo il valore medio teorico (t) che dovremmo aspettarci all'interno di ogni cella se i 4 eventi fossero equiprobabili...	636 / 4 = 159
2.	Calcoliamo lo scarto della frequenza osservata (f) di ogni cella rispetto a quella teorica (t)	125 - 159 = -34
3.	Eleviamo a quadrato questa differenza per diminuire i valori piccoli ed aumentare quelli grandi...	$-34^{22} = 1156$
4.	Dividiamo infine per la frequenza teorica, in modo da standardizzare le distanze...	1156/159 = 7.27
5.	Ripetiamo il procedimento per tutte le celle...	130 ⇒ 5.29 153 ⇒ 0.23 228 ⇒ 29.94
6.	Sommiamo i vari risultati parziali...	42.73

Ci sono in questo procedimento due passaggi (terzo e quarto) che potrebbero essere complessi da capire: il quadrato della differenza rispetto al valore teorico e la sua divisione per il valore teorico. Elevare a quadrato una differenza (tecnica che abbiamo già applicato per il calcolo della varianza), ci permette di ottenere due effetti:

- eliminare il segno negativo;
- amplificare le differenze proporzionalmente alla loro grandezza (il quadrato di 2 è 4, il quadrato di 5 è 25 e quello di 10 è 100).

In questo modo, il numero che otteniamo è tanto più grande quanto maggiore è la differenza di partenza. Dividendo poi questo valore per la frequenza teorica, otteniamo una misurazione che, più o meno, equivale a dire: "quante frequenze teoriche stanno in questo scarto quadratico". Si utilizza quindi ciascuna frequenza teorica come unità di misura per esprimere lo scarto.

In pratica abbiamo costruito un numero che rappresenta *la somma ponderata degli scarti delle frequenze di ciascuna cella rispetto alla sua attesa teorica* (e che è la statistica di χ^2).

E' semplice allora capire come, maggiore è il valore trovato (il χ^2) e maggiore è lo scostamento delle frequenze osservate rispetto a quelle teoriche che ci dovremmo aspettare. Vale a dire, più i dati osservati e quelli teorici sono diversi fra loro (si allontanano), maggiori saranno le loro differenze e più grande diventerà il valore della statistica di χ^2. Più simili i dati teorici a quelli osservati e più piccolo sarà il valore dell'indice statistico finale.

Nel caso che stiamo considerando, maggiore sarà il valore finale e maggiore la probabilità che la distribuzione non sia casuale ma in qualche modo influenzata da qualcosa.

E' altrettanto facile capire come il valore trovato dipenda (per la sua grandezza) anche dal numero di celle e dal numero di frequenze totali: quante più celle possiede la tabella, tanto maggiore sarà la probabilità che una di esse si comporti in modo anomalo; quanto più alto il totale, quanto più è probabile trovare valori elevati della statistica di chi-quadro.

	A_1	A_2	A_3
B_1	60	53	12
B_2	53	23	16
B_3	55	48	20

Tabella 2: Tavola di contingenza fra due ipotetiche variabili categoriali

1.1.2 Indipendenza di due variabili categoriali

Un lavoro analogo possiamo farlo su tabelle di contingenza (ossia tabelle a due entrate) che incrociano le frequenze con cui accadono assieme determinate categorie di due variabili. Ad es. una tabella di contingenza che incrocia i valori delle ipotetiche variabili A e B potrebbe essere come quella che compare in Tab. 2.

Anche in questo caso abbiamo dei valori osservati (quelli della tabella) e possiamo calcolare dei valori teorici (basandoci proprio sul concetto di indipendenza probabilistica).

$$H_0 : P(AB) = P(A) \times P(B)$$

$$H_1 : P(AB) \neq P(A) \times P(B)$$

A questo punto il calcolo della statistica di χ^2 ci dirà quanto si discostano i dati osservati da quelli che abbiamo stimato sotto l'ipotesi di indipendenza. Se il valore sarà basso, realtà e teoria si avvicineranno molto; se il valore sarà alto, si discosteranno.

E se la realtà si avvicina molto alla teoria (χ^2 basso), poiché l'ipotesi teorica è che le due variabili siano indipendenti fra loro, concluderemo che le due variabili non si influenzano reciprocamente. Mentre se il χ^2 è alto, non potremo dire che le variabili sono fra loro indipendenti, ma dovremo affermare che, in qualche modo, si influenzano reciprocamente.

1.1.3 Differenze con un modello teorico

Infine, a partire da una *qualunque configurazione di valori osservati*, e una *qualunque ipotesi teorica*, possiamo applicare la statistica del χ^2 per vedere se l'ipotesi teorica serve per spiegare i dati reali. In questo caso è l'ipotesi alternativa che ipotizza l'equiprobabilità, mentre l'ipotesi nulla fa riferimento ad un modello esplicativo che genera la distribuzione dei dati.

Dal momento che i valori teorici verrebbero stimati sulla base di una teoria, di un modello, di un'ipotesi e la statistica di χ^2 sarebbe tanto più piccola quanto più teoria e realtà sono vicine fra loro. Ovvero, se la teoria *spiega sufficientemente bene* la realtà, il chi-quadro avrà un valore piccolo, se non la spiega abbastanza, avrà un valore elevato.

1.2 Terminologia

Prima di proseguire, poniamo alcune basi terminologiche.

Solitamente è possibile indicare i valori reali di una distribuzione, usando una lettera (generalmente x, y e z) per indicare la variabile e una lettera per indicare un indice (generalmente i, j e k). Usando questo tipo di notazione, possiamo riscrivere la Tab.2 in modo generico, in questo modo:

	A_1	A_2	A_3
B_1	f_{11}	f_{12}	f_{13}
B_2	f_{21}	f_{22}	f_{23}
B_3	f_{31}	f_{32}	f_{33}

	A_1	A_2	A_3	tot.
B_1	f_{ij}	...	...	$f_{i.}$
B_2	...	...	...	...
B_3	...	...	...	...
tot.	$f_{.j}$	...	...	$f_{..}$

Tabella 3: Tabella generica

In questa notazione, usiamo la lettera f per indicare la frequenza di ogni cella e, in particolare, f_{11} indicherà il contenuto della cella all'incrocio fra la riga 1 e la colonna 1, f_{32} la cella all'incrocio fra la riga 3 e la colonna 2; quindi $f_{11} = 60$, $f_{12} = 53$, $f_{31} = 55$.

Possiamo scrivere la stessa tabella in un modo più generico (Tab.).

dove f_{ij} indica le singole celle (al variare degli indici i e j), mentre $f_{i.}$, $f_{.j}$ e $f_{..}$ sono rispettivamente i totali di riga, i totali di colonna e il totale generale[2].

Alcune ovvie relazioni possono essere espresse in termini matematici all'interno di questa tabella, usando le stesse notazioni:

a) il totale della i-esima riga è pari alla somma delle celle di quella riga (con c che indica il numero delle colonne):

$$f_{i.} = \sum_{j=1}^{c} f_{ij}$$

b) analogamente per i totali della colonna j-esima (con r che indica il numero delle righe):

$$f_{.i} = \sum_{i=1}^{r} f_{ij}$$

c) la numerosità totale della tabella corrisponde al totale generale che è pari alla somma di tutti i totali di riga, ovvero alla somma di tutti i totali di colonna oppure alla somma di tutte le celle:

$$N = f_{..} = \sum_{j=1}^{c} f_{.j} = \sum_{i=1}^{r} f_{i.} = \sum_{i=1}^{c} \sum_{j=1}^{r} f_{ij}$$

1.3 La formula di chi-quadro

Se trasformiamo il procedimento di calcolo usato al par. 1.1.1 in formula, poiché abbiamo solo una riga, possiamo scrivere:

$$\chi^2 = \sum_{i=1}^{c} \frac{(f_i - t_i)^2}{t_i}$$

dove indichiamo con c il numero di colonne, i è l'indice che assume tutti i valori fra 1 e c, f_i è il valore della cella (frequenza ottenuta) e t_i è il corrispondente valore teorico.

Se avessimo utilizzato una tabella a doppia entrata (come la Tab. **??**), la formula sarebbe invece:

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(f_{ij} - t_{ij})^2}{t_{ij}}$$

[2] Alcuni testi usano il segno + al posto del punto e quindi f_{i+}, f_{+j} e f_{++}.

in cui i e j indicano rispettivamente le righe (r) e le colonne (c). In genere, però, la formula di chi quadro la si trova scritta in modi più generici, usando notazioni di derivazione anglosassone, ad es.:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} \tag{1}$$

In questo contesto f_o significa frequenza osservata (*observed*) e f_e frequenza teorica o attesa (*expected*) e il simbolo di sommatoria indica di sommare tutti i valori disponibili (ovvero tutte le celle della tabella).

[Formule alternative

$$\chi^2 = \sum \frac{f_o^2}{f_e} - N$$

$$\chi^2 = \sum \frac{N(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)}$$

]

1.4 I valori teorici

Osservando bene la formula 1 vediamo che la parte più importante ma anche "sconosciuta" è proprio la frequenza attesa. In effetti, in base al modo in cui si calcola la frequenza attesa, cambia l'uso che si fa del χ^2.

1.4.1 Equiprobabilità

Al paragrafo 1.1.1 abbiamo calcolato i valori teorici dividendo N per il numero di celle ($t_i = f_i/c$). Questo perché avevamo 1 sola variabile. L'ipotesi sottintesa era che i dati si distribuissero in modo casuale. In pratica abbiamo testato l'ipotesi che le frequenze nelle quattro stagioni fossero equiprobabili.

$$\begin{aligned} H_0: &\quad P(\text{Estate}) = P(\text{Aut}) = P(\text{Inv}) = P(\text{Prim}) = .25 \\ H_1: &\quad P(\text{Estate}) \neq P(\text{Aut}) \neq P(\text{Inv}) \neq P(\text{Prim}) \neq .25 \end{aligned}$$

1.4.2 Indipendenza

Con 2 variabili, le cose cambiano un poco. Vediamo come.

Ipotizziamo di voler verificare se, in un campione di 42 soggetti di entrambi i sessi, la distribuzione dei livelli di educazione sia (H_1) o meno (H_0) dipendente dal sesso (Tab. 4).

Tabella 4: Dati fittizi

		Livello Educativo		
Sesso		Basso	Alto	Totale
	Maschi	13	9	22
	Femmine	13	7	20
Totale		26	16	42

Se precedentemente abbiamo utilizzato il totale generale e lo abbiamo diviso per il numero delle celle (cioè, in questo caso, 42 / 4 = 10.5), adesso dobbiamo considerare che vi sono dei vincoli. Abbiamo solo 20 soggetti di sesso *femminile* e non potremo mai aspettarci, neppure

teoricamente, che siano 21 (cioè 10.5 + 10.5); analogamente abbiamo solo 16 soggetti di livello educativo *Alto* (e non 21). I nostri valori teorici devono quindi tener conto di questi totali che "costringono" i risultati in una certa direzione; per calcolare i valori attesi (per ogni cella) si utilizza una formula abbastanza semplice: si moltiplicano fra loro i totali di riga e di colonna di quella cella e si divide il risultato per il totale generale, in formula:

$$t_{ij} = \frac{f_{i.} f_{.j}}{f_{..}} \tag{2}$$

Applicando questa formula ad ogni elemento della Tab.**??**, avremmo:

		Freq.	Freq. teorica
Maschi	Basso	13	22 x 26 / 42 = 13.62
	Alto	9	22 x 16 / 42 = 8.38
Femmine	Basso	13	20 x 26 / 42 = 12.38
	Alto	7	20 x 16 / 42 = 7.62

La scelta di questo metodo di calcolo non è casuale. Infatti questa semplice formula di calcolo corrisponde alla probabilità teorica che si verifichino contemporaneamente 2 eventi fra di loro indipendenti ovvero la cui probabilità di comparsa di uno dei due non incide sulla probabilità del secondo e viceversa. Dallo studio della probabilità, sappiamo che tale valore è dato dal prodotto delle probabilità dei singoli eventi. Nel nostro caso, per la prima cella della tabella, si tratta di incrociare la probabilità di essere maschio [P(M)] con la probabilità di avere un basso livello educativo [P(B)]:

$$P(MB) = P(M)P(B)$$

La teoria della probabilità frequentista ci dice che la probabilità di un evento è data dalla frequenza con cui compare quell'evento, divisa per il totale degli eventi, quindi:

$$P(M) = \frac{\text{tot. maschi}}{\text{tot. generale}} = \frac{22}{42} = 0.5238$$

$$P(B) = \frac{\text{tot. basso livello}}{\text{tot. generale}} = \frac{26}{42} = 0.6190$$

$$P(MB) = 0.5238 \cdot 0.6190 = 0.3243$$

Poiché abbiamo 42 soggetti, dobbiamo moltiplicare N per la probabilità della prima cella, al fine di stimare quanti soggetti dovrebbero stare in quella cella:

$$42 \cdot 0.3243 = 13.6188$$

Se scriviamo tutto il procedimento in un colpo solo, vediamo facilmente che la formuletta 2 è ricavata da tutto questo ragionamento:

$$N \cdot P(MB) = N \cdot P(M) \cdot P(B) = 42 \cdot \frac{22}{42} \cdot \frac{26}{42} = \frac{22 \cdot 26}{42}$$

1.4.3 Un esempio

Partendo dai dati della Tab. 2, proviamo a calcolare un chi-quadro completo. Iniziamo calcolando i totali di riga e di colonna sulle frequenze osservate (f_o).

	f_o			
	A_1	A_1	A_1	tot.
B_1	60	53	12	125
B_2	53	23	16	92
B_3	55	48	20	123
tot.	168	124	48	34

I valori teorici vengono calcolati come:

f_t		
$125 \times 168/340$	$125 \times 124/340$	$125 \times 48/340$
$92 \times 168/340$	$92 \times 124/340$	$92 \times 48/340$
$123 \times 168/340$	$123 \times 124/340$	$123 \times 48/340$

Notate come, per ogni colonna, vi sia una parte della formula che non cambia (analogamente se consideriamo le formule per riga). Facendo i conti a mano, possiamo semplificarli così:

168 / 340 = 0.49	124 / 340 = 0.36	48 / 340 = 0.14
125×0.49	125×0.36	125×0.14
92×0.49	92×0.36	92×0.14
123×0.49	123×0.36	123×0.14

Nel trascrivere i dati, ho arrotondato a 2 cifre decimali mentre sarebbe opportuno utilizzare tutti i decimali possibili

Otteniamo così le seguenti frequenze teoriche:

f_t		
61.25	45.00	17.50
45.08	33.12	12.88
60.27	44.28	17.22

Possiamo adesso applicare la formula per il calcolo del chi-quadro, dapprima calcolando, per ogni cella, il valore della formula e, successivamente sommando il tutto (volendo essere un esempio dettagliato, farò tutti i passaggi e userò tutti i decimali del visore di una normale calcolatrice).

L'applicazione della formula ad ogni cella:

$(60 - 61.25)^2/61.25$	$(53 - 45)^2/45$	$(12 - 17.5)^2/17.5$
$(53 - 45.08)^2/45.08$	$(23 - 33.12)^2/33.12$	$(16 - 12.88)^2/12.88$
$(55 - 60.27)^2/60.27$	$(48 - 44.28)^2/44.28$	$(20 - 17.22)^2/17.22$

Il calcolo della differenza:

$(-1.25)^2/61.25$	$(8)^2/45$	$(-5.5)^2/17.5$
$(7.92)^2/45.08$	$(-10.12)^2/33.12$	$(3.12)^2/12.88$
$(-5.27)^2/60.27$	$(3.72)^2/44.28$	$(2.78)^2/17.22$

Il quadrato della differenza:

1.5625/61.25	64/45	30.25/17.5
62.7264/45.08	52.6064/33.12	13.69/12.88
27.7729/60.27	9.7344/44.28	7.7284/17.22

La divisione finale:

0.0255102	1.4222222	1.7285714
1.3914463	1.5883575	0.7557764
0.460808	0.3091689	0.4488037

Sommando il contenuto di tutte le celle e arrotondando a due decimali, otteniamo un χ^2 di 8.13.

E ora che abbiamo calcolato la statistica di chi-quadro, cosa ce ne facciamo?

Al paragrafo 1.1 avevamo scritto che la statistica di chi-quadro serviva per stabilire fino a punto potevamo accettare le frequenze ottenute come analoghe, simili, vicine a quelle teoriche e che più alto era il valore trovato, tanto più era *improbabile* che tale lontananza fosse casuale.

Dobbiamo a questo punto fare un procedimento di inferenza statistica.

1.5 La distribuzione chi-quadro

[Forma della curva, cambi al variare dei gl, problemi di numerosità e di freq. teoriche]

1.6 I gradi di libertà

Riprendiamo in considerazione la Tab.2, con i suoi totali.

Il concetto di gradi di libertà nasce dal fatto che avendo 168 eventi nella categoria A1, dovendoli suddividere nelle 3 celle corrispondenti alle categorie di B, noi abbiamo libertà di mettere quanti eventi vogliamo in 2 sole celle... la terza è "costretta" a contenere gli eventi restanti. Lo stesso ragionamento viene fatto per A2, A3 e per ciascuno dei valori di B.

Nella tabella quindi vi sono delle celle (per convenzione le ultime) che non possono contenere "qualsiasi numero" ma solo quanto resta per poter sommare al totale degli eventi di quella categorie. Nella tabella che segue queste celle sono indicate in grassetto.

	A_1	A_2	A_3	tot.
B_1	60	53	**12**	125
B_2	53	23	**16**	92
B_3	**55**	**48**	**20**	123
tot.	168	124	48	34

Il numero delle celle "libere", corrisponde ai gradi di libertà (in inglese, *degree of freedom*, abbreviato in *df*). La formula generale, facilmente comprensibile dall'esempio precedente, è:

$$gl = (r-1)(c-1)$$

ossia numero di righe per numero di colonne, a ciascuno dei quali viene precedentemente sottratto uno.

1.7 L'inferenza

Una volta calcolato il valore finale di un chi-quadro si applica il solito meccanismo del livello di significatività, facendo riferimento alla *distribuzione di chi-quadro* e ai gradi di libertà impli-

cati. Il valore di significatività trovato indica il rischio che noi corriamo, la probabilità che un determinato valore di chi-quadro sia casuale.

Ritornando all'esempio di Tab.1, per sapere se il valore di chi-quadro da noi trovato (42.73) è significativo, consultiamo le tavole relative della distribuzione di chi-quadro (la Tab.5 ne riporta una parte).

Le tavole della distribuzione di chi-quadro riportano, generalmente, i valori critici di chi-quadro per i vari gradi di libertà e per diversi valori α. Cerchiamo quindi la riga corrispondente a 3 gradi di libertà (cioè 4-1) e quindi avanziamo alla ricerca di un valore che sia superiore a quello da noi trovato. Nessuno dei valori segnati sulla riga supera il valore di 42.73, quindi la probabilità a esso connessa è così piccola da essere inferiore allo .001, cioè all'1 per mille.

Tabella 5: Valori critici di chi-quadro (estratto)

	.05	.01	.001
gl=1	3.841	6.635	10.828
2	5.991	9.210	13.816
3	7.815	11.341	16.266
4	9.488	13.277	18.467
5	11.070	15.086	20.515

La maggior parte dei programmi per computer, oltre a fornire il valore di chi-quadro e i gradi di libertà, fornisce anche un valore di probabilità (a volte chiamato "significatività") che permette di interpretare immediatamente il valore statistico calcolato senza utilizzare le tabelle. Se utilizzassimo un programma statistico per rifare lo stesso chi-quadro, otterremmo questi risultati:

	Patologia
Chi-square	42.73
df	3
Sig.	.000

Con 3 gradi di libertà (*df = degree of freedom*), un chi-quadro pari a 42.73 è da considerarsi molto significativo; in effetti la significatività è pari a .000 (che significa che vi è almeno una cifra diversa da zero a partire dal quarto decimale e che tale cifra non viene visualizzata per motivi di arrotondamento) ovvero vi è meno di 1 probabilità su 10.000 che i nostri dati siano così diversi tra loro per puro caso. Nel caso fittizio da noi considerato dovremmo quindi concludere che effettivamente nasce un numero di soggetti schizofrenici diverso rispetto alle stagioni di nascita e in particolare in *inverno*.

Senza bisogno di applicare le formule, si possono consultare delle apposite tavole, che forniscono la probabilità associata ad un certo valore di χ^2, per un dato grado di libertà.

Usando le tavole, dobbiamo seguire il seguente procedimento:

1. fissiamo un livello α (di solito α =.05);
2. calcoliamo i gradi di libertà;
3. troviamo sulle tavole la riga corrispondente ai gradi di libertà e la scorriamo fino alla colonna corrispondente al livello α;
4. all'incrocio fra riga e colonna, troviamo il valore critico di χ^2 ($\chi^2{}_c$);
5. se il nostro chi-quadro è inferiore al valore critico, accettiamo l'ipotesi H_0;
6. se è superiore, accettiamo l'ipotesi alternativa.

Esempio:
χ^2 =10.63; gl=4; α =.05

$\chi^2 < \chi^2_c$	accetto H_0
$\chi^2 > \chi^2_c$	accetto H_1

Tabella 6: Regole di accettazione/rifiuto

Consultando le tavole, trovo un χ^2 critico di 9.48
Poiché 10.63 è maggiore del valore critico, rifiuto H_0 e accetto H_1

1.8 Correzione di Yates

Riprendendo, invece, i dati di Tab.3, già confrontando ad occhio i valori teorici e i valori ottenuti, possiamo aspettarci che il chi-quadro non sia significativo, poiché i due valori sono molto vicini fra loro. In effetti, se calcoliamo la statistica con un programma per computer, otteniamo:

Pearson Chi-square	.155
Continuity Correction	.006
Df	1
Sig.	.694

Il risultato del chi-quadro (*Pearson Chi-Square*) è pari a 0.155 che, con 1 grado di libertà non risulta significativo: la probabilità esatta calcolata (*Sig.*) è pari infatti a 0.694, cioè: se decidessimo di accettare l'ipotesi H_1 che il sesso influisce sul livello economico di un individuo, correremmo un rischio di sbagliare del 69%; rischio che è considerato eccessivo e che ci induce ad accettare l'ipotesi opposta.

Poiché la tabella è composta da 4 celle in forma 2x2, viene calcolato un altro indice di chi-quadro (*Continuity correction*), conosciuto anche come "correzione di Yates", che permette di adeguare maggiormente la distribuzione del chi-quadro di una tabella 2x2 alla curva di chi-quadro.

1.9 Verifica di un modello

Fino ad ora abbiamo utilizzato la statistica di chi-quadro per verificare se una determinata distribuzione era (oppure no) uniformemente distribuita. Per questo motivo, abbiamo calcolato i valori delle frequenze teoriche come rapporti ponderati delle righe e delle colonne e, per accettare l'ipotesi H_1, ci aspettavamo di trovare valori di chi-quadro molto elevati e statisticamente associati ad un basso valore di α.

Ma poiché la tecnica del chi-quadro confronta una distribuzione realmente ottenuta con una teorica, noi possiamo utilizzare questo test anche per verificare un nostro particolare modello di dati. In questo caso, però, un valore elevato di chi-quadro (quindi significativo), vorrebbe dire che la distribuzione reale dei nostri dati si discosta dalla distribuzione teorica che ci aspettavamo mentre un valore bassissimo o nullo, significherebbe che la nostra teoria spiega bene i dati da noi trovati.

Come esempio usiamo quello iniziale. Leggiamo un articolo (ipotetico) in cui si afferma che nel periodo invernale (rispetto alle altre stagioni) nascono più soggetti che poi riveleranno disturbi di tipo schizofrenico. L'autore dell'articolo precisa anche che in genere, durante l'inverno, nella sua popolazione di riferimento, sono nati circa il 55% di tutti i soggetti con tali disturbi. Noi allora prendiamo i dati in nostro possesso e calcoliamo un normale test di chi-quadro, che ci risulta significativo. A questo punto ci chiediamo se le caratteristiche del nostro campione sono simili a quelle del campione dell'autore dell'articolo. Ricalcoliamo il chi-quadro, usando questa

volta come frequenze teoriche i valori che ricaveremo dai dati dell'articolo, ad es. pari al 37% per l'inverno e al 18% per la primavera, al 20% per l'estate e al 25% per l'autunno (Tab.7). Il χ^2 così ottenuto è pari a 1,48, che per 1 grado di libertà non è statisticamente significativo.

Tabella 7: Confronto con un articolo di riferimento (Dati fittizi)

	Primavera	Estate	Autunno	Inverno	Totale
freq. osservate	125	130	153	153	636
% di riferimento	18	20	25	37	
freq. attese	114,48	127,2	159	235,32	636

Tuttavia, in questa circostanza, un chi-quadro *non significativo*, indica che la distribuzione dei dati da noi ottenuta è "vicina" a quella teorica ipotizzata e che quindi il nostro campione è simile a quello utilizzato nell'articolo di riferimento.

1.10 Riepilogo

Ora che abbiamo spiegato anche a livello intuitivo la statistica di chi-quadro, affrontiamo brevemente i criteri da considerare nella sua applicazione.

- Può essere usata con una o due variabili categoriali o ordinali;
- L'attribuzione di un caso (soggetto) ad una categoria/cella dev'essere univoca, ovvero un soggetto classificato in una cella non deve comparire contemporaneamente in un'altra: questo si chiama *indipendenza dei casi*;
- Non si può applicare (è rischioso) se più del 20% delle celle ha una "frequenza attesa" inferiore a 5 (solo nel caso di una tabella 2x2, si può utilizzare una formula alternativa chiamata "correzione di Yates"); oppure se una cella ha frequenza attesa inferiore a 1. In questo caso, se vi sono 3 o più categorie e se il "significato" logico di tali categorie lo permette, è possibile accorparne alcune in modo da ampliare la numerosità di quella particolare riga/colonna. Ad es. è possibile far confluire la categoria "convivente" con quella di "sposato" e la categoria "vedovo" con "divorziato" se ciò che importa nell'analisi è l'ampiezza del nucleo familiare;
- Quando il numero di celle è piccolo e il numero di casi è grande conviene "verificare" la validità del chi-quadro tramite l'uso del coefficiente C di contingenza.
- Quando si usa il chi-quadro su tabelle di contingenza con più di 2 righe o colonne, e si trova un valore significativo di χ^2, si vorrebbe anche sapere quale cella o quali celle sono responsabili della significatività. Questa conoscenza aiuta molto nell'interpretare i risultati del test. Esistono delle tecniche abbastanza complesse, chiamate "tecniche di partizione" che permettono di andare a vedere come si comportano le celle o alcuni gruppi di celle rispetto a tutte le altre. Tralasciando queste tecniche di partizione, suggeriamo l'uso dei residui standardizzati, proposti da Haberman (1973). Quando il residuo standardizzato di una cella supera il valore di 2 (in realtà 1.96, pari ad $\alpha = .05$), la cella si discosta dal suo valore teorico a sufficienza per essere considerata come una cella anomala, che ha contribuito a rendere significativo il test di chi-quadro.

Se dovete calcolare un chi-quadro su dati già in forma tabellare, anziché usare un complesso programma statistico, è più semplice usare un programma apposito. Nel mondo di internet ve ne sono due facilmente utilizzabili: il primo è un programma in italiano per il sistema operativo Dos[3], mentre il secondo programma è in inglese (`chi1`), più completo, ed è disponibile in

[3] `http://web.newsguy.com/germano/soft/chiquadro.php`

qualunque *mirror* di SimtelNet nella directory di statistica[4].

1.11 Fonti

Blalock jr. (1960, pp. 345-360), Cristante, Lis, e Sambin (1982, pp. 57-72).

Riferimenti bibliografici

Blalock jr., H. M. (1960). *Social statistics*. New York: McGraw-Hill Book. (Trad. it. *Statistica per la ricerca sociale*. Milano: Il Mulino, 1969.)

Cristante, F., Lis, A., & Sambin, M. (1982). *Statistica per psicologi*. Firenze: Giunti Barbèra.

[4] ad es. in `http://sunsite.cnlab-switch.ch/ftp/mirror/simtelnet/msdos/statstcs/chi1_0.zip`