Analisi di dati qualitativi

Psicometria 1 - Lezione 13
Lucidi presentati a lezione

AA 2000/2001 dott. Corrado Caudek

Alcuni esperimenti non producono risultati misurabili su una scala continua ma bensì *dati categoriali.*

In precedenza è stata descritta la distribuzione binomiale risutante da un esperimento consistente in *n* prove, ciascuna delle quali può produrre soltanto due esiti possibili. Frequentemente si incontrano situazioni analoghe in cui però il numero degli esiti possibili è maggiore di due.

Uno psicologo sociale, ad esempio, potrebbe studiare le reazioni al comportamento autoritario e usare una classificazione in *k* categorie.

Gli esperimenti che producono un numero discreto di esiti qualitativamente diversi sono detti *multinomiali* e costituiscono una generalizzazione di un esperimento binomiale.

Un esperimento multinomiale ha le seguenti proprietà:

1. L'esperimento è costituito da n prove aventi caratteristiche equivalenti.
2. L'esito di ciascuna prova viene classificato in una di k categorie.
3. La probabilità che l'esito di una singola prova cada nella i-esima categoria, con $i = 1, 2, \ldots, k$, è p_i e rimane costante in tutte le prove. Si noti che $p_1 + p_2 + \ldots + p_k = 1$.
4. Le prove sono tra loro indipendenti.
5. Le variabili di interesse sono $n_1, n_2, \ldots, n_k$, laddove n_i, con $i = 1, 2, \ldots, k$, è il numero di prove classificate nell'i-esima categoria. Si noti che $n_1 + n_2 + \ldots + n_k = n$.

Supponiamo di eseguire un esperimento che produce un insieme finito di risultati. Supponiamo inoltre di disporre di un modello teorico che predice la distribuzione dei risultati dell'esperimento.

Se l'esperimento viene ripetuto molte di volte, diventa possibile verificare l'adeguatezza delle predizioni teoriche nei confronti dei dati empiricamente ottenuti.

Sia X una variabile aleatoria che rappresenta i possibili risultati dell'esperimento considerato e sia $m(x)$ la distribuzione di probabilità di X.

L'adeguatezza di un modello teorico di predire i risultati di un esperimento empirico si può valutare calcolando la statistica seguente:

$$V = \sum_j \frac{\left(o_x - n \cdot m(x)\right)^2}{n \cdot m(x)}$$

dove, per ciascuno dei possibili risultati dell'esperimento, o_x denota la frequenza osservata, $m(x)$ denota la probabilità teorica di quel risultato (così come ipotizzato dal modello) e n è il numero di osservazioni.

Il prodotto $n\ m(x)$ rappresenta dunque la frequenza attesa per una classe di osservazioni, ovvero la frequenza predetta dal modello.

$$V = \sum \frac{(f_o - f_t)^2}{f_t}$$

La distribuzione campionaria della statistica V è conosciuta.

Per moderati e grandi valori di n, la statistica V segue approssimativamente la distribuzione c^2 con $\nu = n - 1$ gradi di libertà.

Esempio. Supponiamo di eseguire l'esperimento consistente nell'estrarre una pallina da un'urna contenente 50 palline colorate. Le palline sono rosse, bianche, verdi, blu e nere.

L'ipotesi che vogliamo sottoporre a verifica è che i cinque colori sono equamente rappresentati nell'urna. In base a quest'ipotesi, dunque, la distribuzione teorica è $p = 1/5$ per ciascuno degli esiti possibili (pallina estratta rossa, bianca, verde, blu o nera).

Se l'esperimento venisse ripetuto 100 volte (con reimmissione), la frequenza attesa di ciascuno dei cinque esiti possibili sarebbe $f_i = 100/5 = 20$.

Supponiamo di eseguire l'esperimento descritto e di ottenere i seguenti dati:

Colore della pallina estratta	Frequenza osservata
rosso	25
bianco	5
verde	16
blu	30
nero	24

La statistica *V* diventa:

$$V = \sum \frac{(f_o - f_t)^2}{f_t}$$

$$V = \frac{(25-20)^2}{20} + \frac{(5-20)^2}{20} + \frac{(16-20)^2}{20} + \frac{(30-20)^2}{20} + \frac{(24-20)^2}{20} = 19.1$$

Come si valuta la significatività della statistica osservata? Dobbiamo usare un test ad una o due code?

L'ipotesi nulla dice che la differenza tra le frequenze osservate e quelle attese è attribuibile soltanto all'errore di campionamento. Se non ci fosse errore di campionamento, le frequenze osservate sarebbero uguali alle frequenze attese e la statistiva *V* sarebbe uguale a zero.

In che circostanze viene contraddetta l'ipotesi nulla?

Questo succede quando le frequenze osservate sono molto diverse dalle frequenze attese.

Dato che la differenza tra frequenze osservate e frequenze attese è elevata al quadrato, maggiori sono le discrepanze tra i dati predetti e quelli osservati, maggiore sarà il valore assunto dalla statistica *V*. La regione critica sarà dunque quella corrispondente alla coda destra della distribuzione.

Con $\alpha = .05$ e $\nu = 4$ il valore critico di χ^2 è 9.48773.

NIntegrate[((x)^((n-2)/ 2)) Exp[-x/ 2] / (2^(n/ 2) Gamma[n/ 2]),{x, 9.48773, Infinity}] = 0.05

Il valore osservato di 19.1 cade nella regione di rifiuto.

Possiamo quindi rigettare l'ipotesi nulla e concludere che i quattro colori non sono rappresentati nell'urna nella stessa misura.

Esempio. Uno sperimentatore studia le abitudini legate al fumo. In particolare si chiede se ci sono stati dei cambiamenti nelle abitudini relative al fumo negli ultimi 10 anni.

Lo sperimentatore dispone di un campione casuale di 863 fumatori di età compresa tra i 40 e i 50 anni. Gli individui di questo campione vengono classificati in base al numero di pacchetti di sigarette che vengono fumate quotidianamente.

Categoria (# di pacchetti di sigarette)	Frequenza osservata
0	406
1	164
2	189
3	78
4 o più	26
	863

Un indagine equivalente condotta 10 anni fa aveva prodotto i seguenti risultati:

Categoria	Frequenza relativa
0	.43
1	.17
2	.24
3	.10
4 o più	.06
	1.0

Il problema è quello di stabilire se le abitudini relative al fumo sono cambiate oppure no nel corso degli ultimi 10 anni.

Sia $\alpha = .01$.

Nella tabella successiva sono riportate le frequenze osservate insieme alle frequenze attese calcolate in base all'ipotesi che *la popolazione abbia esattamente la stessa distribuzione di 10 anni fa*.

Le frequenze attese sono calcolate come: $n\ m(x)$.

Categoria	Frequenza osservata	Frequenza attesa
0	406	371.09
1	164	146.71
2	189	207.12
3	78	86.30
4 o più	26	51.78
	863	863

$$V = \frac{(406-371.09)^2}{371.09} + \frac{(164-146.71)^2}{146.71} + \frac{(189-207.12)^2}{207.12}$$
$$+ \frac{(78-86.3)^2}{86.3} + \frac{(26-51.78)^2}{51.78} = 20.54$$

Quale è il limite critico che delimitano la regione di rifiuto?

Con $\alpha = .01$ e 4 gradi di libertà, il valore critico è 13.2767.

NIntegrate[((x)^((n-2)/ 2)) Exp[-x/ 2] / (2^(n/ 2) Gamma[n/ 2]),{x,13.2767,Infinity}] = 0.01

Il valore osservato cade all'interno della regione di rifiuto. Se l'ipotesi nulla fosse vera (ovvero, se il comportamento relativo al fumo fosse lo stesso di 10 anni fa), la statistica *V* assumerebbe un valore maggiore o uguale a quello osservato nel campione *in meno dell'uno per cento dei casi*.

Il ricercatore può dunque rigettare l'ipotesi nulla e concludere che le abitudini relative al fumo sono cambiate negli ultimi 10 anni.

TABELLE DI CONTINGENZA

Consideriamo due variabili qualitative *R* e *C*.

Supponiamo che la variabile *R* sia definita in base ad un numero di categorie mutuamente esclusive ed esaustive uguale a *J*, e la variabile *C* definisca un numero di categorie mutuamente esclusive ed esaustive pari a *K*.

Costruiamo una tabella nelle cui celle sono riportate le frequenze che indicano quanto spesso una data combinazione delle due variabili si è verificata nei dati.

La tabella avrà un numero di celle uguale a $J * K$ e la frequenza di ciascuna cella sarà indicata con f_{jk}.

La somma delle frequenze di tutte le celle sarà uguale a *n* (il numero totale dei casi del campione).

Questo tipo di tabella si chiama *tabella di contingenza.*

	C_1	C_2	...	C_k	...	C_K	
R_1	f_{11}	f_{12}	...	f_{1k}	...	f_{1K}	$f_{1.}$
R_2	f_{21}	f_{22}	...	f_{2k}	...	f_{2K}	$f_{2.}$
...	...	...	...	...	...	...	...
R_j	f_{j1}	f_{j2}	...	f_{jk}	...	f_{jK}	$f_{j.}$
...	...	...	...	...	...	...	...
R_J	f_{J1}	f_{J2}	...	f_{Jk}	...	f_{JK}	$f_{J.}$
	$f_{.1}$	$f_{.2}$	...	$f_{.k}$	...	$f_{.K}$	n

La frequenza f_{jk} è la frequenza della cella formata dalla riga R_j e dalla colonna C_k.

La frequenza marginale della riga R_j è $f_{j.}$ e la frequenza marginale della colonna C_k è $f_{.k}$

Per sottoporre a verifica l'ipotesi di indipendenza tra le variabili R e C consideriamo innanzitutto p_j, ovvero la probabilità che un'osservazione cada nella j-esima categoria del criterio di classificazione R.

Se i criteri di classificazione R e C sono indipendenti, una stima di p_j viene data dalla frequenza marginale osservata nella j-esima riga divisa per il numero totale di osservazioni:

$$\hat{p}_j = \frac{f_{j.}}{n}$$

In maniera analoga, una stima della probabilità che un'osservazione cada nella categoria C_k sarà data da

$$\hat{p}_k = \frac{f_{.k}}{n}$$

Se ***l'ipotesi nulla di indipendenza tra i due criteri di classificazione R e C fosse vera***, una stima della probabilità attesa in ciascuna cella della tabella di contingenza, p_{jk}, verrebbe fornita dal prodotto tra le probabilità marginali (stimate) della *j*-esima riga e della *k*-esima colonna:

$$\hat{p}_{jk} = \hat{p}_{j.} \times \hat{p}_{.k} = \frac{n_{j.}}{n} \times \frac{n_{.k}}{n}$$

In base all'ipotesi nulla di indipendenza tra i due criteri di classificazione *R* e *C*, dunque, una stima della ***frequenza attesa*** in ciascuna cella della tabella di contingenza, $E(f_{jk})$, può essere ottenuta in base al prodotto tra le frequenze marginali della *j*-esima riga e della *k*-esima colonna diviso per la grandezza totale del campione:

$$\hat{E}\left(f_{jk}\right) = n\left(\hat{p}_{j.} \cdot \hat{p}_{.k}\right) = n\left(\frac{n_{j.}}{n}\right)\left(\frac{n_{.k}}{n}\right) = \frac{n_{j.} \cdot n_{.k}}{n}$$

Una volta stimate le frequenze attese, l'ipotesi di indipendenza tra i criteri di classificazione R e C viene sottoposta a verifica calcolando la statistica V e valutando questa statistica in base alla distribuzione c^2.

$$V = \sum \frac{(f_o - f_t)^2}{f_t}$$

Quanti sono i gradi di libertà?

Il numero appropriato di gradi di libertà per valutare la statistica V in base alla distribuzione c^2 è uguale al numero delle celle della tabella di contingenza meno 1 grado di libertà per ciascuno dei vincoli lineari indipendenti che vengono imposti sulle frequenze osservate.

Una tabella di contingenza è composta da $J' K$ celle.

Da questo numero deve essere sottratto 1 grado di libertà a causa del vincolo

$$f_{1.} + f_{2.} + \ldots + f_{j.} + \ldots + f_{J.} = n$$

e 1 grado di libertà a causa del vincolo

$$f_{.1} + f_{.2} + \ldots + f_{.k} + \ldots + f_{.K} = n$$

Per una tabella di contingenza $J \times K$, dunque, i gradi di libertà saranno dunque uguali a:

$$n = (J - 1)(K - 1)$$

Esempio. Supponiamo di volere confrontare 4 programmi di recupero scolastico. 797 studenti hanno partecipato ai diversi programmi e vengono classificati in 3 categorie: quelli che hanno passato il test dopo il programma, quelli che hanno fallito il test una volta dopo il completamento del programma e quelli che hanno fallito il test più di una volta dopo il completamento del programma.

Si sottoponga a verifica l'ipotesi nulla di assenza di associazione tra il tipo di programma di recupero scolastico e le prestazioni nel test.

Sia $\alpha = .01$.

Programma	B_1 (nessun fallimento)	B_2 (un fallimento)	B_3 (più di un fallimento)	*fj.*
A_1	*122*	*70*	*8*	200
	(115.68)	*(73.27)*	*(11.04)*	
A_2	*141*	*39*	*15*	195
	(112.79)	*(71.44)*	*(10.77)*	
A_3	*106*	*79*	*18*	203
	(117.42)	*(74.73)*	*(11.21)*	
A_4	*92*	*104*	*3*	199
	(115.11)	*(72.91)*	*(10.99)*	
f.k	461	292	44	797

Le cifre tra parentesi rappresentano le *stime delle frequenze attese* in base all'ipotesi di indipendenza:

$$\hat{f}_{jk} = \frac{f_{j.} \times f_{.k}}{n}$$

Ad esempio, per la cella della prima riga e della prima colonna avremo:

$$\hat{f}_{11} = (200 \cdot 461) / 797 = 115.68$$

Per la presente tabella di contingenza, la statistica *V* diventa:

$$V = \frac{(122-115.68)^2}{115.68} + \frac{(70-73.27)^2}{73.27} + \ldots + \frac{(3-10.99)^2}{10.99} = 54.0$$

e si distribuisce come c^2 con (4 - 1) (3 - 1) = 6 gradi di libertà.

Quale è il valore che delimita la regione di rifiuto con 6 gradi di libertà?

n=6
NIntegrate[((x)^((n-2)/ 2)) Exp[-x/ 2] / (2^(n/ 2) Gamma[n/ 2]),
{x, 16.8119, Infinity}] = 0.00999998

Dato che il valore osservato di c^2 cade all'interno della regione di rifiuto, possiamo rifiutare l'ipotesi nulla. Concludiamo dunque che c'è un'associazione significativa tra il tipo di programma di recupero e le prestazioni nel test: le prestazioni degli studenti dipendono dal tipo di programma di recupero che è stato seguito.

Spesso è utile esaminare le frequenze delle singole celle per individuare le deviazioni maggiori dalle frequenze attese.

Se *n* è grande, per ciascuna cella possiamo calcolare il *residuo standardizzato*:

$$z_{jk} = \left(f_{osservata} - f_{attesa} \right) / \sqrt{f_{attesa}}$$

Questi residui standardizzati possono essere considerati come variabili aleatorie distribuite normalmente in base all'ipotesi di assenza di associazione.

L'esame dei residui standardizzati ci consente di stabilire quali sono le celle che forniscono il contributo maggiore all'associazione rivelata dal test del c^2.

Inoltre, il segno della differenza dalla frequenza attesa indica se l'associazione è positiva (si osserva un valore maggiore di quello atteso) o negativa (si osserva un valore minore di quello atteso) in una particolare combinazione di categorie di righe e colonne.

Se α = .05 allora, in base alla correzione di Bonferroni, il livello di significatività per ciascun residuo sarà:

.05 / 12 = .004

(dove 12 è il numero di celle nella tabella).

Con un test bidirezionale, dunque, un residuo può essere considerato significativo soltanto se $|z|$ è maggiore di 2.86.

```
1-NIntegrate[(1/ Sqrt[2 Pi]) Exp[-(x^2)/ 2],{x,-2.86, 2.86}]
= 0.00423641
```

Nella riga A_2 e colonna B_2, ad esempio, il residuo standardizzato si calcola come:

$$z_{22} = (39 - 71.44) / \sqrt{71.44} = -3.84$$

Questo valore eccede (in valore assoluto) il valore critico di 2.86.

I residui standardizzati nelle altre celle vengono calcolati e valutati allo stesso modo.

In conclusione, i dati esaminati suggeriscono che il programma A_2 è quello con la maggiore capacità di fare in modo che gli studenti non incorrano in ulteriori fallimenti nel test del corso in questione.

Per i programmi A_1 e A_3 troviamo delle frequenze osservate che non si discostano significativamente da quelle attese. Per il programma A_4 troviamo delle frequenze significativamente superiori a quelle attese *nella direzione indesiderata.*

Caso speciale di una tabella 2x2

Nel caso di un tabella 2x2, il calcolo della statistica *V* (ovvero, il χ^2 di Pearson) è molto semplice.

Consideriamo la tabella seguente:

a	**b**	**a+b**
c	**d**	**c+d**
a+c	**b+d**	**n**

Le lettere *a*, *b*, *c*, *d* rappresentano le frequenze in ciascuna cella.

In questo caso, *V* si calcola come:

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

Nel caso di una tabella 2x2, il valore della statistica *V* è solitamente corretto per ottenere un'approssimazione migliore alla distribuzione χ^2 (con un grado di libertà):

$$c^2 = \frac{n(|ad - bc| - N/2)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Questa è la *correzione di Yates* per la continuità e si applica soltanto quando c'è un solo grado di libertà.

ASSUNZIONI

Il χ^2 è uno dei test statistici più semplici e viene usato spesso.

Tuttavia, la facilità con la quale è possibile calcolare questa statistica non deve farci dimenticare che l'uso del χ^2 è appropriato solo se vengono rispettate alcune condizioni.

1. Il test del χ^2 è basato sull'assunzione che le osservazioni sono indipendenti. Questo significa che non si può applicare il test del χ^2 nel caso in cui vi può essere qualche forma di dipendenza tra le osservazioni considerate. Per esempio, questo spesso avviene quando delle osservazioni ripetute sono fatte con i medesimi individui. Non è il fatto che le osservazioni siano ripetute che ovviamente produce una forma di dipendenza tra le osservazioni. Questo dipende dalla natura dell'esperimento e dal tipo di dati. Per esempio, un esperimento condotto per mezzo di osservazioni ripetute sugli stessi soggetti potrebbe creare una qualche forma di apprendimento del compito sperimentale e, dunque, in questo modo l'assunzione di indipendenza verrebbe certamente violata.

2. Ciascuna delle osservazioni che vengono categorizzate deve poter qualificarsi per una e soltanto per una cella della tabella di contingenza.

3. Il problema più grande che dobbiamo affrontare quando applichiamo il test del χ^2 riguarda la grandezza del campione.

La statistica *V* si distribuisce come χ^2 ***soltanto quando la grandezza del campione è infinita***. Se il campione è troppo piccolo, dunque, l'approssimazione risulta essere inadeguata.

La regola che si usa per stabilire sele dimensioni di un campione sono adeguate è la seguente:

Per tabelle con più di un singolo grado di libertà, la frequenza attesa in ciascuna cella deve essere di almeno 5.

Per tabelle con un singolo grado di libertà, la frequenza attesa in ciascuna cella deve essere di almeno 10.

Questa regola è molto severa e in certi casi frequenze attese minori sono accettabili, in particolare quando il numero di gradi di libertà è molto grande.

Per un numero elevato di gradi di libertà è accettabile di usare il test del χ^2 anche se, in alcune (poche) celle della tabella, vi sono frequenze attese con un valore di 1.

ESERCIZIO

Cento pazienti schizofrenici sono stati divisi in due gruppi in base alla gravità dei sintomi dimostrati. Il gruppo "schizofrenia grave" era costituito da 80 pazienti, mentre quello "schizofrenia lieve" era costituito da 20 pazienti. Il "disadattamento sociale" di questi pazienti è stato misurato usando una scala da 1 a 100 e i pazienti sono stati divisi in due gruppi: quelli fortemente disadattati (60 pazienti) e quelli senza problemi di adattamento sociale (40 pazienti). Sappiamo che 50 pazienti fortemente disadattati dimostrano anche sintomi gravi. Lo psicologo ipotizza che vi sia un'associazione tra gravità dei sintomi e disadattamento sociale. In base a questi dati, che cosa potete concludere?