

Statistica

per le scienze
del comportamento



Joan Welkowitz
Barry Cohen
Robert Ewen

Chi quadrato e bontà d'adattamento: il caso di una variabile

Consideriamo il seguente problema. Un editore ha tre proposte da valutare per il titolo di un nuovo settimanale letterario per donne – *Domani*, *Scelta* e *Nuove Alternative* – e si chiede se le tre proposte siano egualmente popolari tra il pubblico femminile. Per verificare questa ipotesi seleziona un campione casuale di 177 donne e chiede loro di esprimere una preferenza per uno dei tre titoli. I risultati indicano 65 preferenze per *Domani*, 60 per *Scelta* e 52 per *Nuove Alternative*: questi risultati permettono di rifiutare l'ipotesi nulla per cui, *nella popolazione*, un ugual numero di donne preferisce ciascuno dei tre titoli? Oppure questi risultati hanno una probabilità abbastanza alta di verificarsi a causa dell'errore di campionamento quando l'ipotesi nulla è vera, portandoci alla conclusione che i tre titoli raccolgono la stessa popolarità nel pubblico femminile?

Nel capitolo precedente abbiamo visto come verificare l'ipotesi nulla quando un evento discreto ha due soli possibili esiti. In questo capitolo generalizzeremo tali procedure a tutte quelle situazioni in cui l'esito possibile di un certo evento cade in una tra tre o più categorie. Detto in un altro modo, vedremo come verificare un'ipotesi quando i dati sono espressi in termini di *occorrenze di più di due categorie di eventi*; in questo caso, ciò che faremo sarà confrontare le frequenze osservate con delle frequenze teoriche calcolate sulla base di una qualche ipotesi.

I dati del problema esposto all'inizio del paragrafo sono espressi in termini di frequenza o di conteggio – quante persone preferiscono una certa alternativa – e il nostro obiettivo è decidere se sia legittimo concludere che le frequenze osservate seguono una specifica distribuzione. Nell'esempio, l'editore sa quante donne preferiscono ogni titolo nel campione considerato; è però interessato a sapere se, da quei dati, può concludere che nell'intera popolazione le preferenze per i titoli raccolgono lo stesso numero di donne.

La strategia per verificare questo tipo di ipotesi è simile a quella per dati dicotomici di cui ci siamo occupati nel capitolo precedente. Comunque, non possiamo usare la distribuzione binomiale per calcolare la probabilità associata a successioni di eventi con più di due esiti distinti. Per tre categorie dovremmo riferirci alla distribuzione trinomiale; più in generale, il modello statistico adatto a un qualsiasi numero di esiti possibili è costituito dalla *distribuzione multinomiale*.

Così come la distribuzione normale può essere usata per approssimare i dati ottenuti in modo più preciso dalla distribuzione binomiale, una statistica opportuna può essere utilizzata per dati categorici con più di due categorie: la distribuzione di questa statistica può essere ragionevolmente approssimata mediante la distribuzione chi quadrato (ciò non è più vero quando trattiamo frequenze molto basse; ci occuperemo di questo problema più avanti nel capitolo). Questa statistica viene chiamato *chi quadrato* – come la distribuzione che (approssimativamente) segue – e viene rappresentata dalla lettera greca minuscola *chi* elevata al quadrato: χ^2 .

Dovendo trattare in questo capitolo risultati con più di due categorie, non possiamo più usare X per rappresentare la frequenza di una categoria ed $N - X$ per rappresentare la frequenza dell'altra; dovremo invece considerare la frequenza di ciascuna categoria (f_o , da *Observed Frequency*). Allo stesso modo, non avremo più a che fare con un solo valore atteso (quello di X) sulla base dell'ipotesi nulla (e

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

dove

f_o = frequenze osservate

f_e = frequenze attese

$\sum$ viene calcolata su tutte le categorie

Se le differenze tra le frequenze osservate e quelle attese sono piccole, anche χ^2 sarà piccolo; al contrario, maggiori sono le differenze tra le frequenze osservate e quelle attese sulla base di H_o , maggiore sarà χ^2 . Se le differenze complessive tra i valori osservati e quelli attesi sono così grandi da verificarsi per effetto del caso solo o in meno del 5% dei casi quando l'ipotesi nulla è vera, allora rifiuteremo l'ipotesi nulla.

Il problema illustrato all'inizio del capitolo si riferisce alla scelta tra diverse categorie (o livelli) di una *singola* variabile: nell'esempio che abbiamo usato, la variabile è la preferenza per un certo titolo e ogni soggetto sceglie fra tre diverse opzioni.

L'analisi di un problema di questo tipo richiede la costruzione di una tabella che includa le tre possibili scelte con le loro frequenze osservate (Tabella 20.1). Quindi dobbiamo determinare le frequenze attese per poter calcolare il χ^2 . L'ipotesi nulla stabilisce che i tre titoli sono ugualmente graditi al pubblico femminile: se fosse vera, dunque, ci aspetteremo che il campione delle 177 donne intervistate si divida esattamente in tre categorie ugualmente rappresentate. Ne deriva che le frequenze attese sono 177/3, cioè 59, 59 e 59 (osservate che la somma delle frequenze attese deve essere pari alla somma delle frequenze osservate). La domanda ora diventa: quanto è probabile che le frequenze osservate risultino pari rispettivamente a 65, 60 e 52 se le preferenze per i tre titoli sono esattamente le stesse nella popolazione? In altri termini, le frequenze osservate 65, 60 e 52 sono significativamente diverse dalle frequenze attese 59, 59 e 59? La risposta a questa domanda è nel calcolo del χ^2 riportata in Tabella 20.1. Notate come

Tabella 20.1
Test chi quadrato per la scelta del titolo di un nuovo settimanale letterario per donne

Titolo	Frequenza osservata (f_o)	Frequenza attesa (f_e)	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
Domani	65	59	6	36	0,610
Scelta	60	59	1	1	0,017
Nuove alternative	52	59	-7	49	0,831
$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 1,458$					

ciascun valore di $(f_o - f_e)^2$ sia stato diviso per la relativa frequenza attesa; le frequenze attese sono tutte uguali tra loro in questo esempio, ma non sarà sempre così in quanto le frequenze teoriche dipendono dall'ipotesi nulla. I tre valori che risultano da queste divisioni vengono poi sommati tra loro per fornire il valore sperimentale del χ^2 , pari a 1,46.

Per verificare la significatività del χ^2 appena calcolato con uno specifico criterio di significatività, dobbiamo fare riferimento alla Tavola H riportata in Appendice. Osservate come questa tavola specifichi un diverso valore critico del χ^2 per ciascun valore di df ; il chi quadrato, come t , fornisce infatti una famiglia di curve, la cui forma varia a seconda dei gradi di libertà (Figura 20.1). Per la distribuzione del χ^2 il numero di gradi di libertà dipende dal numero delle *categorie*, non dalla numerosità del campione come nel caso di t . Nel caso di una sola variabile, $df = k - 1$, dove k è dato dal numero di categorie della variabile. Nell'esempio qui considerato, quindi, $df = 3 - 1 = 2$; ne discende che, per essere significativo per $\alpha = 0,05$, il χ^2 deve essere uguale o maggiore di 5,99. Dato che il valore calcolato nell'esempio è pari a 1,46, l'ipotesi nulla deve essere accettata: non esiste evidenza sufficiente per negare l'ipotesi per cui le scelte dei tre titoli siano ugualmente frequenti nella popolazione. Non abbiamo dunque dati sufficienti per concludere che uno (due) dei tre titoli sia (siano) preferito(i) agli altri.