ϵ - Tonsille e *Streptococcus pyogenes*

Guido Masarotto
Statistica (metodi di base)

23 maggio 2000

Verifica dell'ipotesi di indipendenza in una tabella a doppia entrata

I dati campionari

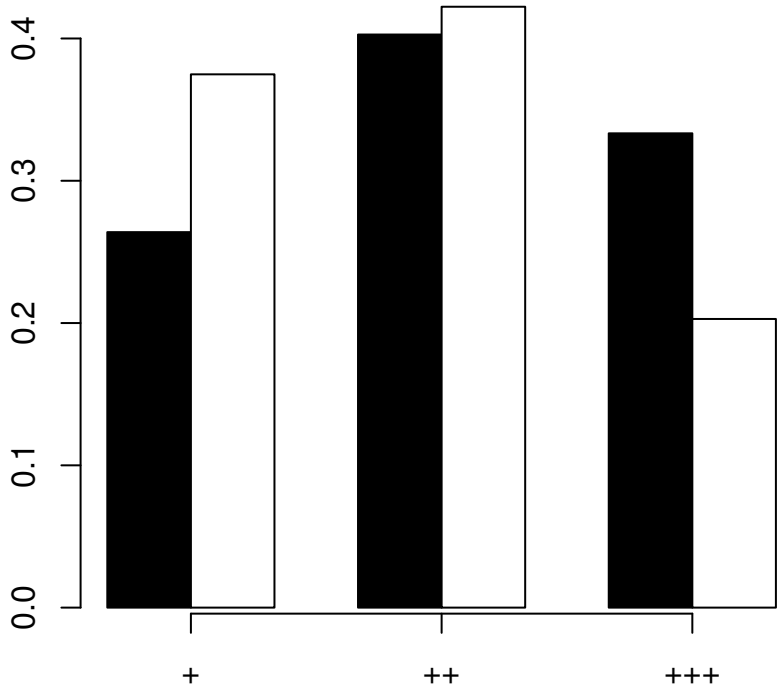
Nel corso di uno studio sulla determinazione di possibili fattori prognostici (predittivi) per alcune malattia otorino-laringoiatriche su 1398 bimbi o ragazzi sono state rilevate le seguenti due variabili:
(a) Presenza (in un tampone nasale) di `Streptococcus pyogenes`; variabile dicotomica con modalità "portatore" e "non portatore".
(b) Stato delle tonsille rilevato utilizzando la scala qualitativa ordinata: (i) normali (abbreviato in $+$), (ii) leggermente ingrossate ($++$) e (iii) ingrossate ($+++$).
I bimbi erano stati scelti casualmente tra tutti gli individui tra i 3 e i 15 di età residenti in un ampia e popolosa regione inglese.

Il problema che affrontiamo è se esiste o no una qualche forma di associazione tra le due variabili.

La seguente tabella mostra le frequenze assolute.

Streptococcus	Tonsille			
pyogenes	$+$	$++$	$+++$	Totale
portatore	19	29	24	72
non portatore	497	560	269	1326
Totale	516	589	293	1298

Grafico a barre



Distribuzione di “stato delle tonsille” condizionata a “portatore” (barre nere) e “non portatore” (barre bianche). L’altezza della barre è proporzionale alle frequenze relative. I portatori sembrano avere le tonsille “più grosse”.

Frequenze attese e X^2 di Pearson

La seguente tabella mostra le frequenze attese nell'ipotesi di perfetta indipendenza in distribuzione.

Streptococcus pyogenes	Tonsille $+$	$++$	$+++$	Totale
portatore	$26{,}6$	$30{,}3$	$15{,}1$	72
non portatore	$489{,}4$	$558{,}7$	$277{,}9$	1326
Totale	516	589	293	1298

Il valore della statistica X^2 di Pearson è $7{,}88$.

Si osservi che, rispetto alla tabella attesa, nella tabella osservata ci sono troppi portatori con tonsille ingrossata e troppo pochi portatori con tonsille normali. E che il viceversa accade per i non portatori.

La popolazione di riferimento

• L'analisi precedente mostra che la distribuzione di "stato delle tonsille" è diversa tra i portatori e i non portatori. Quindi, nella tabella di pagina ϵ-1 non esiste indipendenza in distribuzione.

• Una domanda che è spontaneo porsi è se e a chi è possibile estendere questi risultati. In realtà, se ci pensa questa è la vera domanda. Infatti, ci scusino i 1398 ragazzi, ma le tonsille di alcuni sconosciuti, probabilmente, non sono uno dei nostri principali problemi. I dati, viceversa, ci possono interessare per quello che ci possono raccontare sulla relazione intercorrente $in\ generale$ tra `Streptococcos Pyogenes` e tonsille.

• Gli elementi del campione sono stati estratti casualmente tra i bimbi di una particolare regione geografica. Possiamo allora pensare che ci possano parlare direttamente della relazione esistente tra le due variabili in questo più grande gruppo di individui. Ovvero, l'insieme dei bimbi e ragazzini tra 3 e 15 abitanti nella regione inglese considerata costituisce quella che usualmente viene chiamata la **popolazione di riferimento**. Quello che vogliamo fare è "interrogare" i dati campionari per ottenere informazioni sulle caratteristiche di questa popolazione.

Breve digressione sui bimbi norvegesi, italiani, nigeriani,. . .

• Sarebbe interessante se i dati ci parlassero di tutti i bambini del mondo.

• Però questo richiede che non ci siano differenze, rispetto ai caratteri considerati, tra i bimbi inglesi (anzi di una particolare regione dell'Inghilterra) e, ad esempio, i bimbi nigeriani. Infatti nel campione non ci sono bimbi nigeriani. E quindi, tutto quello di particolare che riguarda quest'ultimi non può essere studiato con questi dati.

• Ovvero, un campione di bimbi inglesi è al più **rappresentativo** di tutti i bimbi inglesi (ovvero della popolazione da cui è stato estratto)[1]. Noi possiamo anche decidere che le conclusioni che i dati ci suggeriscono valgono anche per i bimbi della Nigeria. Ma si tratta appunto di una $nostra$ decisione. E, come è ovvio, estendere le conclusioni di una indagine su di una popolazione ad altre popolazioni è intrisincamente $pericoloso$. L'estensione può avvenire solo tramite nuovi studi (sulle altre popolazioni). Fino a che questi non sono condotti, le conclusioni su di una popolazione sono, al più, ipotesi da verificare per le altre.

[1] E può anche non esserlo se l'estrazione è stata in qualche forma truccata (si pensi al giudice Ford dell'unità δ)

Ascensori, aspirine e la mutabilità dei comportamenti umani

• Quanto detto deve *sempre* essere tenuto presente. In modo particolare, per gli studi nell'ambito delle scienze sociali.

• I meccanismi fisici, chimici e biologici sono piuttosto stabili nel tempo e nello spazio. Le leggi con cui si costruiscono gli ascensori a Oslo e a Sidney sono le stesse. E in tutte le farmacie del mondo contro il mal di testa si trovano prodotti che contengono acido acetilsalicilico (il prodotto commerciale più comune è l'aspirina). E, sempre senza differenza tra razze e ambienti, l'abuso di acido acetilsalicilico aumenta il rischio di gastrite.

• Lo stesso non si può dire per i fenomeni sociali. Due comunità separate da pochi chilometri possono avere comportamenti molto diversi. La stessa comunità a distanza di pochi anni può presentare comportamenti diversi,. . .

Una tabella *fantasma*

• Ritorniamo a considerare l'insieme dei bimbi tra i 3 e i 15 anni residenti nella regione considerata.

• Se le due variabili fossere state rilevate su *tutti* i bimbi avremmo potuto costruire una tabella, analoga a quella di pagina ϵ-1, del tipo

Streptococcus	Tonsille			
pyogenes	$+$	$++$	$+++$	Totale
portatore	F_{11}	F_{12}	F_{13}	F_{1+}
non portatore	F_{21}	F_{22}	F_{23}	F_{2+}
Totale	F_{+1}	F_{+2}	F_{+3}	N

dove (i) N indica il numero di bimbi in quell'area dell'Inghilterra, (ii) F_{11} il numero di bimbi che sono portatori ma hanno le tonsille normali, (iii). . .

• La tabella precedente noi non la conosciamo. Per questo è una tabella $fantasma$ e per questo le varie frequenze sono state indicate con lettere.

Che relazione esiste tra la tabella osservata e quella *fantasma*?

• Il campione è stato formato: (1) estraendo un bimbo tra gli N della popolazione; (2) estraendo un altro bimbo tra gli $N-1$ bimbi non estratti alla prima estrazione; . . . ; (1398) Estraendo un bimbo tra gli $N-1397$ bimbi non estratti nelle prime 1397 estrazioni. In tutte le estrazioni, è stata assegnata probabilità uguale a tutti i bimbi non ancora estratti.

• Dividiamo tutte le frequenze della tabella $fantasma$ per N ottenendo

Streptococcus	Tonsille			
pyogenes	$+$	$++$	$+++$	Totale
portatore	π_{11}	π_{12}	π_{13}	π_{1+}
non portatore	π_{21}	π_{22}	π_{23}	π_{2+}
Totale	π_{+1}	π_{+2}	π_{+3}	1

• Vista la maniera con cui è stato formato il campione, la probabilità che il primo bimbo sia, ad esempio, un non portatore di *Streptococcus pyogenes* con le tonsille normali è π_{21}. Le successive estrazioni non sono tra di loro indipendenti. Infatti, escludere i bimbi già estratti altera ovviamente l'urna da cui stiamo estraendo.

• Nel caso in esame però N è molto grande e quindi la dipendenza è trascurabile da un punto di vista pratico.

• Quindi, almeno approssimativamente, la tabella di pagina ϵ-1 mostra come si sono ripartiti nelle 6 "categorie" (portatori,+), (portatori,++), . . . , (non portatori,+ + +) i risultati di 1398 esperimenti casuali indipendenti tutti caratterizzati da

$$\begin{cases} \text{P(estrarre un (portatore,+))} = \pi_{11} \\ \text{P(estrarre un (portatore,++))} = \pi_{12} \\ \qquad\qquad\qquad\qquad \vdots \\ \text{P(estrarre un (non portatore,+++))} = \pi_{23} \end{cases} .$$

• Ma allora

$$(F_{11}, F_{12}, \ldots, F_{23}) \sim \text{Multinomiale}(n, (\pi_{11}, \pi_{12}, \ldots, \pi_{23}))$$

dove $(F_{11}, F_{12}, \ldots, F_{23})$ indica la variabile casuale che descrive il numero di (portatori,+), (portatori,++) e così via estratti[2].

[2]Si ricordi che il racconto di una variabile casuale multinomiale in termini di palline colorate e di urne è: se (i) esiste un urna contenente palline di k colori diversi; (ii) tutte le palline possono essere estratte con la stessa probabilità; (iii) la frazione di palline del colore i-simo è π_i (ad esempio, se l'i-simo colore è "viola" allora $\pi_i = 0.12$ indica che il 12% delle palline dell'urna è "viola"); (iv) n palline sono estratte dall'urna con *reintroduzione* (ovvero la composizione dell'urna non cambia) allora la variabile casuale $(N_1, \ldots, N_k)$ che descrive il numero di palline estratte del primo colore,del secondo colore,. . . , è una Multinomiale$(n, (\pi_1, \ldots, \pi_k))$.

Verifica dell'ipotesi di indipendenza

Una domanda interessante che possiamo fare ai dati è: nella tabella $fantasma$ esiste indipendenza in distribuzione? In altre parole la dipendenza che abbiamo rilevato nel campione è una peculiarità dei soli bimbi estratti e quindi l'abbiamo osservata per puro caso oppure è la "manifestazione" di una reale associazione tra i due fenomeni esistente nella popolazione.

Si tratta, ovviamente, di un problema di verifica d'ipotesi che può essere scritto nella forma

$$\begin{cases} \mathsf{H}_0 : \pi_{ij} = \pi_{i+}\pi_{+j}, \quad i = 1,2 \;\; j = 1,2,3 \\ \mathsf{H}_1 : \text{le } \pi_{ij} \text{ non rispettano i vincoli previsti da } \mathsf{H}_0 \end{cases}$$

La statistica test più usata è l'X^2 di Pearson. E' certamente una statistica appropriata visto che ci aspettiamo che assuma valori (i) piccoli quando H_0 è vera e (ii) grandi quando è falsa.

La distribuzione approssimata di X^2

E' possibile mostrare (rinviamo al solito la dimostrazione di questo risultato a corsi più avanzati) che se H_0 è vera e nessuna frequenza attesa è troppo piccola allora la distribuzione di X^2 può essere approssimata con la distribuzione di una variabile casuale chiamata χ^2 di Pearson.

La distribuzione χ^2 dipende da un solo parametro, chiamato i gradi di libertà della distribuzione, che nel caso che stiamo trattando (verifica dell'ipotesi di indipendenza in una tabella di contingenza) deve essere posto uguale a

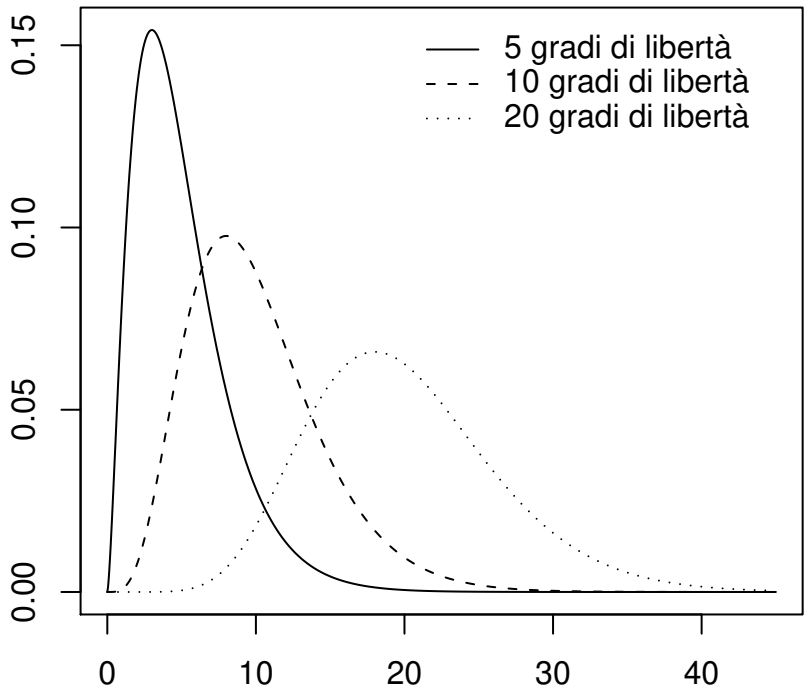
$$\left[\begin{pmatrix}\text{numero righe}\\ \text{tabella}\end{pmatrix} - 1\right] \times \left[\begin{pmatrix}\text{numero colonne}\\ \text{tabella}\end{pmatrix} - 1\right]$$

Ad esempio, per la tabella in esame, i gradi di libertà sono $2 = (2-1) \times (3-1)$.

L'approssimazione è ritenuta "decorosa" se la più piccola delle frequenze attese (si noti, quelle attese, non quelle osservate) è più grande di 1 e migliora man mano che queste aumentano.

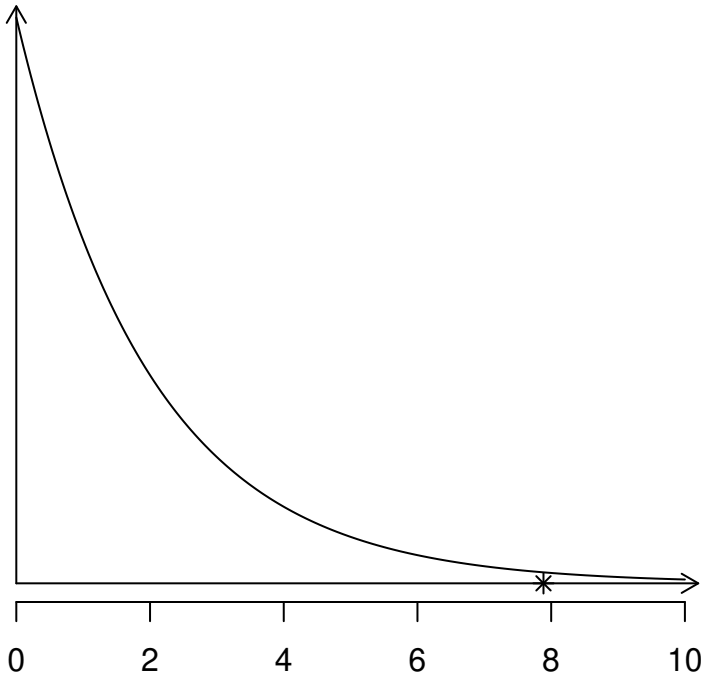
Alla fine di questa unità è allegata una tabella con pre-calcolati alcuni percentili di un χ^2.

Densità di una variabile casuale χ^2 per tre valori dei gradi di libertà



Si osservi che all'aumentare dei gradi di libertà la densità si sposta verso destra. Infatti, è possibile dimostrare che la media della variabile casuale coincide con i gradi di libertà.

Test: analisi grafica del risultato



Densità di una v.c. χ^2 con 2 gradi di libertà. L'asterisco sull'asse delle ascisse segna il valore osservato della statistica test. Il valore è "moderatamente" ma non "esageratamente" spostato verso destra, ovvero, verso H_1. La conclusione potrebbe essere una sorta di "dubbioso rifiuto di H_0".

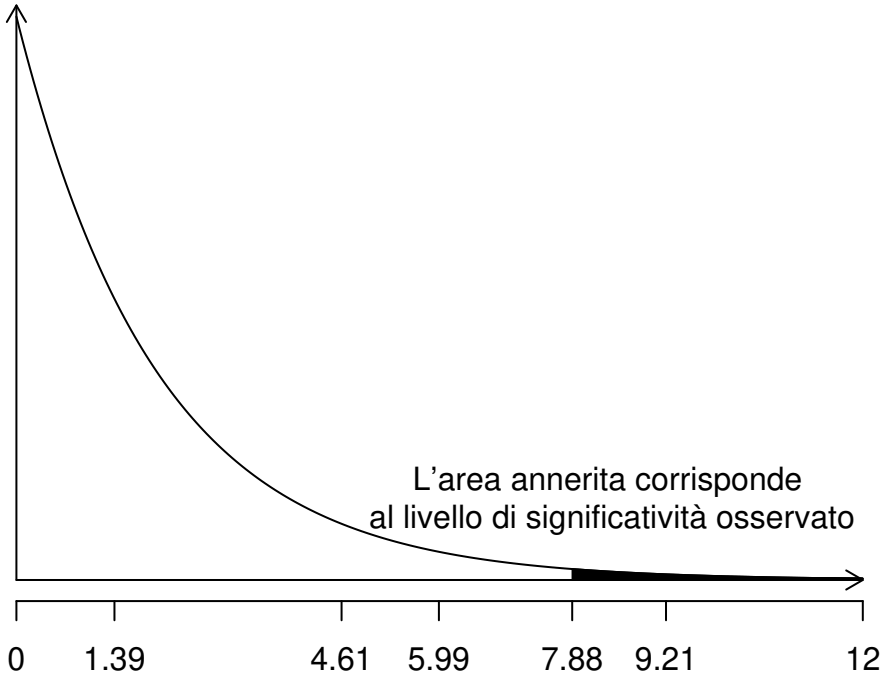
Livello di significatività osservato (e suo calcolo approssimato da una tavola dei percentili)

• "Lontano da H_0" vuol dire per il test che stiamo considerando "grande". Quindi, in questo caso il livello di significatività osservato è la probabilità, di osservare quando è vera H_0 un valore uguale o maggiore di quello osservato. Quindi, per i dati presentati in questa unità,

$$\begin{pmatrix}\text{livello}\\ \text{significatività}\\ \text{osservato}\end{pmatrix} = P(\chi^2 \text{ con } 2 \text{ gradi libertà} \geq 7{,}88)$$

• Supponiamo ora di voler determinare un intervallo che lo contenga conoscendo solo alcuni percentili della distribuzione. Ad esempio, supponiamo di conoscere solamente la seguente tabella in cui $\chi^2_{2,p}$ indica il percentile p-simo di un χ^2 con 2 gradi di libertà.

p	0,5	0,90	0,95	0,99
$\chi^2_{2,p}$	1,39	4,61	5,99	9,21



• Il valore del test $(7{,}88)$ è compreso tra il 95-simo e il 99-simo percentile. Ora, per definizione la probabilità di assumere un valore più grande del 95-simo (99-simo) percentile è 5% (1%). Perciò

$$0{,}01 \leq \text{(livello significatività osservato)} \leq 0{,}05 \qquad (1)$$

Esercizi

1. Si costruisca una regola del tipo (accetto H_0)/(rifiuto H_0) che accetti H_0 quando H_0 è vera con una probabilità prefissata, diciamo $1 - \alpha$. La si applichi ai dati considerati usando sia $\alpha = 0{,}05$ che $\alpha = 0{,}01$.
2. Si supponga di voler calcolare un intervallo di confidenza per la proporzione di portatori dello `Streptococcus pyogenes` (diciamo che deve contenere il vero valore con probabilità $0{,}9$). Come potremmo fare?.
3. Lo studente si convinca che la (1) e il grafico di pagina ϵ-13 forniscono essenzialmente la stessa informazione.

Quantili di un χ^2 di Pearson

g indica i gradi di libertà, p la probabilità lasciata a "sinistra", Quindi, ad esempio, $P(\chi^2$ con 2 gradi di libertà $\leq 9{,}21) = 0{,}99$

	p							
g	0,1	0,25	0,5	0,75	0,90	0,95	0,99	0,999
1	0,02	0,1	0,45	1,32	2,71	3,84	6,63	10,83
2	0,21	0,58	1,39	2,77	4,61	5,99	9,21	13,82
3	0,58	1,21	2,37	4,11	6,25	7,81	11,34	16,27
4	1,06	1,92	3,36	5,39	7,78	9,49	13,28	18,47
5	1,61	2,67	4,35	6,63	9,24	11,07	15,09	20,52
6	2,2	3,45	5,35	7,84	10,64	12,59	16,81	22,46
7	2,83	4,25	6,35	9,04	12,02	14,07	18,48	24,32
8	3,49	5,07	7,34	10,22	13,36	15,51	20,09	26,12
9	4,17	5,9	8,34	11,39	14,68	16,92	21,67	27,88
10	4,87	6,74	9,34	12,55	15,99	18,31	23,21	29,59
11	5,58	7,58	10,34	13,7	17,28	19,68	24,72	31,26
12	6,3	8,44	11,34	14,85	18,55	21,03	26,22	32,91
13	7,04	9,3	12,34	15,98	19,81	22,36	27,69	34,53
14	7,79	10,17	13,34	17,12	21,06	23,68	29,14	36,12
15	8,55	11,04	14,34	18,25	22,31	25	30,58	37,7
16	9,31	11,91	15,34	19,37	23,54	26,3	32	39,25
17	10,09	12,79	16,34	20,49	24,77	27,59	33,41	40,79
18	10,86	13,68	17,34	21,6	25,99	28,87	34,81	42,31
19	11,65	14,56	18,34	22,72	27,2	30,14	36,19	43,82
20	12,44	15,45	19,34	23,83	28,41	31,41	37,57	45,31
21	13,24	16,34	20,34	24,93	29,62	32,67	38,93	46,8
22	14,04	17,24	21,34	26,04	30,81	33,92	40,29	48,27
23	14,85	18,14	22,34	27,14	32,01	35,17	41,64	49,73
24	15,66	19,04	23,34	28,24	33,2	36,42	42,98	51,18
25	16,47	19,94	24,34	29,34	34,38	37,65	44,31	52,62
26	17,29	20,84	25,34	30,43	35,56	38,89	45,64	54,05
27	18,11	21,75	26,34	31,53	36,74	40,11	46,96	55,48
28	18,94	22,66	27,34	32,62	37,92	41,34	48,28	56,89
29	19,77	23,57	28,34	33,71	39,09	42,56	49,59	58,3
30	20,6	24,48	29,34	34,8	40,26	43,77	50,89	59,7