LEZIONI DI STATISTICA MEDICA

Prof. Roberto de Marco

Lezione n.13

- Analisi di dati qualitativi

-test per l'adattamento

-test per l'analisi di tabelle di contingenza

Sezione di Epidemiologia & Statistica Medica
Università degli Studi di Verona

TEST PER L'ADATTAMENTO

SCOPO: verificare se una distribuzione empirica si adatta a una qualsiasi distribuzione teorica

Esercizio

I seguenti dati riportano il numero di aborti spontanei in un campione di 70 donne con 4 gravidanze:

n° aborti	0	1	2	3	4
OSSERVATI (O)	24	28	7	5	6

⇨ **N° totale di aborti= 81**

Siamo interessati a sapere se esiste una "suscettibilità" individuale delle donne per l'aborto

Se un precedente aborto non condiziona l'esito di una successiva gravidanza (eventi indipendenti), il numero di aborti è una V.C. con distribuzione binomiale con n=4 e π stimato da:

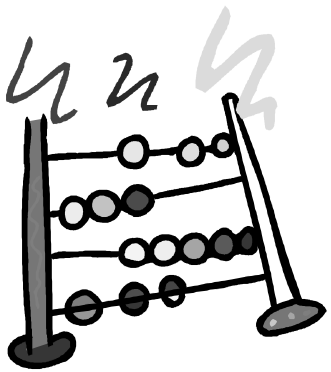
$$p=n^\circ \text{ aborti}/n^\circ \text{nascite}= (81/70\cdot 4)$$

Gli attesi del modello saranno dati da:

$$P(n^\circ \text{ aborti}=x)\cdot 70, \text{ per } 0\leq x\leq 4$$

Attesi (E)

$P(n^\circ \text{ aborti}=0)=0.2551\cdot 70= 17.86$
$P(n^\circ \text{ aborti}=1)=0.4154\cdot 70= 29.08$
$P(n^\circ \text{ aborti}=2)=0.2536\cdot 70= 17.75$
$P(n^\circ \text{ aborti}=3)=0.0688\cdot 70= 4.82$
$P(n^\circ \text{ aborti}=4)=0.0070\cdot 70= 0.49$



n°aborti	0	1	2	3	4
OSSERVATI (O)	24	28	7	5	6
ATTESI (E)	17.86	29.08	17.75	4.82	0.49
(O-E)	**6.14**	**-1.08**	**-10.75**	**0.18**	**5.51**

A questo punto il problema è verificare se lo scarto tra i casi osservati e gli attesi è superiore a quanto ci si possa aspettare per soli motivi casuali.

La statistica che misura il grado di "disaccordo" tra distribuzione osservata e attesa è data da:

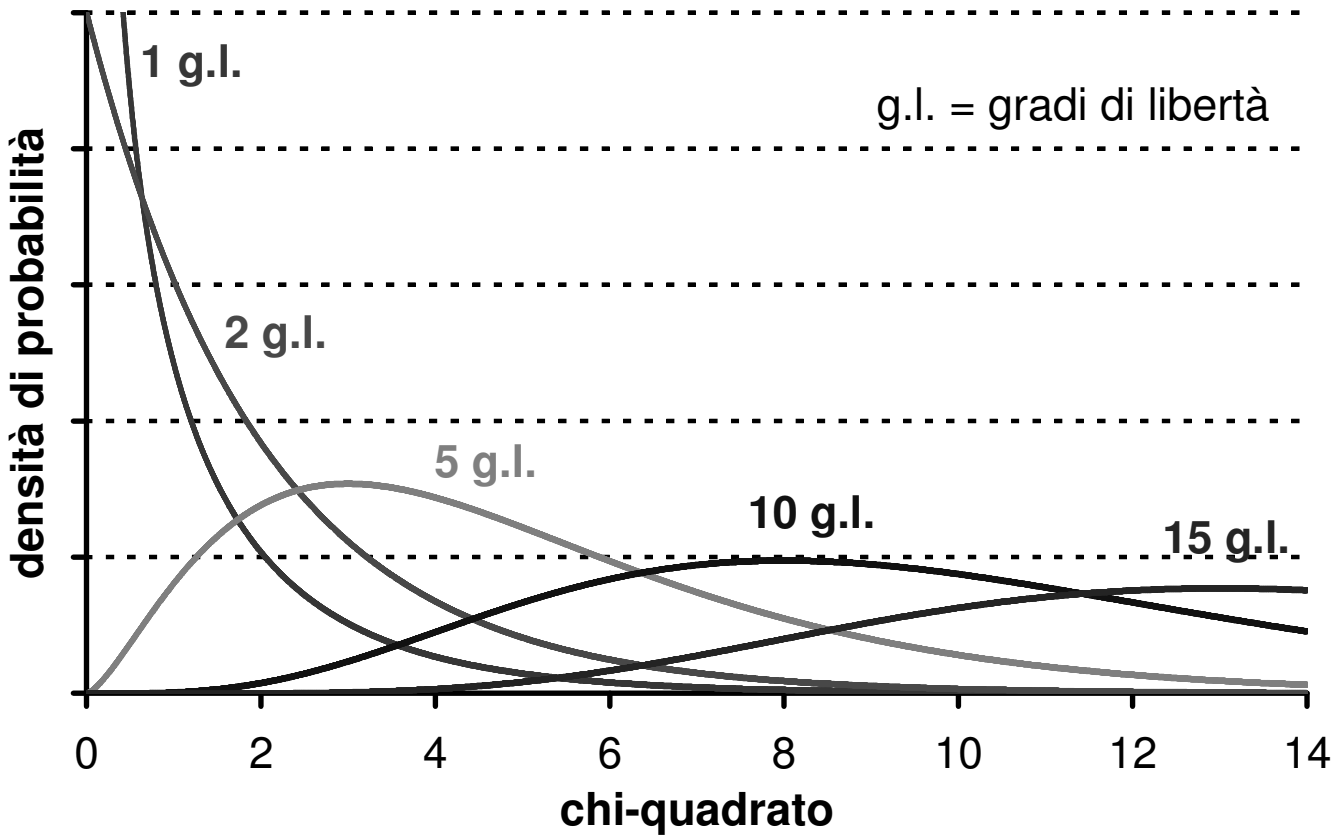
$$\chi^2 = \sum_{i=1}^{k} (O_i - E_i)^2 / E_i$$

Dove:

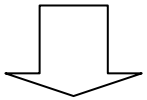
- k=n° di celle (modalità/classi della variabile)
- Oi= frequenze osservate nell'i-ma cella
- Ei=frequenze attese nell'i-ma cella

Sotto l'ipotesi che osservati e attesi provengano dalla stessa distribuzione, la statistica χ^2 ha una distribuizione χ^2 che dipende dal numero di gradi di libertà (g.l.)

g.l.=n° celle-n°parametri stimati per calcolare gli attesi -1



L'utilizzo delle tavole è del tutto simile a quello visto per la distribuzione t di student (l'unica differenza è dovuta al fatto che χ^2 è sempre positivo)



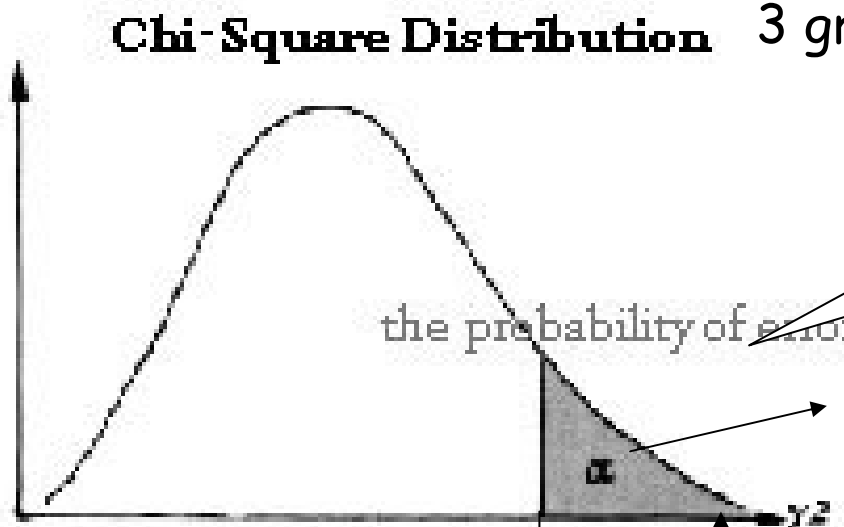
Se il χ^2 è superiore a quello tabulato (per α=0.05/0.001 e competenti g.l.) si rifiuta l'ipotesi che osservati e attesi abbiano la stessa distribuzione, altrimenti non si hanno elementi per rifiutare H_0

g.d.l= n° celle-n°parametri stimati per calcolare gli attesi -1=

= 5(n° celle)-1(n° parametri: p)-1=3

$$\chi^2 = \sum_{i=1}^{5} (O_i - E_i)^2 / E_i =$$

=[(6.14)²/17.86+(-1.08)²/29.08 +(-10.75)²/17.75
+(0.18)²/4.82+(5.51)²/0.49]=70.63



3 gradi di libertà

IL TEST DEL CHI-QUADRATO E' SEMPRE A UNA CODA

α=0.001

16.27< 70.63

Si esclude che l'evento aborto sia indipendente (con errore di I° tipo<0.001);

⇒ i dati sembrano quindi individuare una suscettibilità individuale per l'aborto spontaneo

In genere, se E<5 si aggregano gli attesi e gli osservati delle celle vicine per ottenere una statistica più stabile:

n°aborti	0	1	2	3-4
OSSERVATI (O)	24	28	7	11
ATTESI (E)	17.86	29.08	17.75	5.31

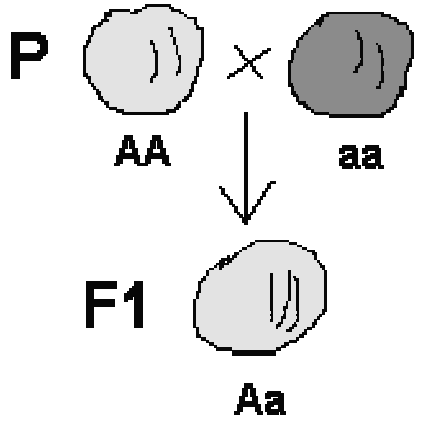
g.d.l= 4-2=2

Esercizio 2

Consideriamo i dati di Mendel sulla segregazione indipendente dei caratteri forma (L=liscia, r=rugosa) e colore (G=giallo, v=verde) del seme di pisello. Mendel reincrociò eterozigoti di prima generazione di genotipo LrGv e fenotipo giallo liscio e ottenne i seguenti risultati:

GL	315
Gr	101
vL	108
vr	32
totale	556

Questi dati confermano la teoria mendeliana della segregazione indipendente dei due caratteri?



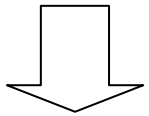
In base all'indipendenza dei caratteri:

- P(giallo-liscio) =9/16
- P(giallo-rugoso)=3/16
- P(verde-liscio) =3/16
- P(verde-rugoso)=1/16

	GL	Gr	vL	vr	totale
osservati (O)	315	101	108	32	556
attesi (E)	312.75	104.25	104.25	34.75	556.00
(O-E)	2.25	-3.25	3.75	-2.75	
(O-E)^2	5.06	10.56	14.06	7.56	37.25
(O-E)^2/E	0.02	0.10	0.13	0.22	0.47

- $\chi^2 = 0.47$
- gdl=4-1=3
- $\chi^2_{3,\,0.05} = 7.82$

0.47<7.82: Non si rifiuta H_0

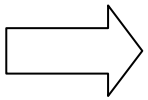


I caratteri "caratteristiche della superficie" e "colore" si segregano indipendentemente l'uno dall'altro (**III legge di Mendel**)

TEST PER L'ANALISI DI TABELLE DI CONTINGENZA

Tabella di contingenza: **tabella che riporta la frequenza associata alla combinazione di due variabili qualitative**

Paziente	sesso	stato di vita
1	M	V
2	F	M
3	M	M
4	F	V
5	M	M
6	M	V
7	F	V
8	F	M
9	M	V
10	M	V



STATO	SESSO: M	SESSO: F	tot
V	4	2	6
M	2	2	4
tot	6	4	10

Nomenclatura di una tabella di contingenza

X

1.....i.....k

Y

1	O_{11}		O_{i1}		O_{k1}
.					
.					
j	O_{1j}		O_{ij}		O_{kj}
.					
t	O_{1t}		O_{it}		O_{kt}

Marginali di riga

$$O_{\cdot j} = \sum_{i=1}^{k} O_{ij}$$

Marginali di colonna

$$O_{i\cdot} = \sum_{j=1}^{t} O_{i\cdot}$$

$$N = \sum_{i=1}^{k} \sum_{j=1}^{t} O_{ij}$$

⇒ Sotto l'ipotesi di indipendenza la probabilità di una qualsiasi combinazione delle modalità delle due variabili è data da:

$$P_{ij} = P_{i.} \cdot P_{.j}$$

Probabilità congiunta

Prodotto delle marginali

⇒ Gli <u>attesi</u> nella cella ij-ma, sotto l'ipotesi di indipendenza, saranno quindi

$$E_{ij} = P_{i.} \cdot P_{.j} \cdot N$$

N° totale di soggetti

Probabilità congiunta

Prodotto delle marginali

$P_{i.}$ viene stimato da $O_{i.}/N$

$P_{.j}$ viene stimato da $O_{.j}/N$

$E_{ij}=(O_{i.}/N)\,(O_{.j}/N)\cdot N=(O_{i.}\cdot O_{.j})/N$

Statistica test:

SI USA SOLO PER IL CONFRONTO TRA FREQUENZE, NON PER PROPORZIONI, TASSI….

$$\chi^2=\sum_{i=1}^{k}\sum_{j=1}^{t}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}}$$

Misura la discrepanza tra dati osservati e l'ipotesi di indipendenza

GRADI DI LIBERTA':

Sotto H_0 : $\chi^2 \sim \chi^2_{(c-1)(r-1)}$

Cioè, sotto l'ipotesi di indipendenza

c= n° di colonne

r= n° di righe

Esercizio 3

In un clinical trial vengono sperimentati due farmaci, A e B: al termine dello studio si valuta l'effetto del farmaco in base allo stato di vita (vivo (V) o morto (M)) del paziente.

I risultati al termine dello studio sono i seguenti:

Tmt A: 41 pazienti deceduti su 257

Tmt B: 64 pazienti deceduti su 244

Qual è il farmaco più efficace?

Soluzione1 :

- Il campione è molto numeroso: per verificare quale trattamento sia migliore applico il test z:
  - p(A)=0.1595, p(B)=0.2663 E.S.=0.0364
  - $z = (p_A - p_B)/ES(p_A - p_B) = -2.81$

I dati possono anche essere rappresentati mediante una tabella di contingenza

	stato di vita		
farmaco	**vivo**	**morto**	
A	216	41	257
B	180	64	244
	396	105	**501**

L'uso dei farmaci influenza lo stato di vita?

Stato di vita e trattamento sono indipendenti?

OSSERVATI

	stato di vita		
farmaco	**vivo**	**morto**	
A	203.14	53.86	257
B	192.86	51.14	244
	396	105	**501**

E_{11}=257*396/501=203.14
E_{12}=257*105/501=53.86
E_{21}=244*396/501=192.86
E_{22}=244*105/501=51.14

ATTESI

$\chi^2=(216-203.14)^2/203.14+(41-53.86)^2/53.86+(180-192.86)^2/192.86+(64-51.14)^2/51.14=\mathbf{7.97}$

Sotto H_0 : $\chi^2 \sim \chi^2_{(2-1)(2-1)=1}$ $\chi^2_{1;0.05}=3.84$

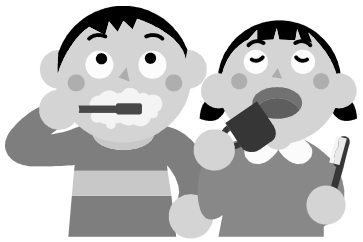
7.97>3.84: Usare un trattamento o l'altro influisce sullo stato di vita dei pazienti.

Esercizio 4

Si riportano di seguito i dati relativi all'igiene orale per tipo di scuola frequentata. Si può affermare che il livello di igiene orale non cambia al variare del tipo di scuola?

	igiene orale				Totale
tipo di scuola	**buona**	**sufficiente**	**insufficiente**	**pessima**	
sotto la media	62	103	57	11	233
media	50	36	26	7.00	119
sopra la media	80	69	18	2.00	169
totale	192	208	101	20	**521**

Soluzione1 :



- Tabella di contingenza nell'ipotesi di indipendenza:

ATTESI

	igiene orale				Totale
tipo di scuola	**buona**	**sufficiente**	**insufficiente**	**pessima**	
sotto la media	85.9	93.0	45.2	8.9	233
media	43.9	47.5	23.1	4.6	119
sopra la media	62.3	67.5	32.8	6.5	169
totale	192	208	101	20	**521**

$\chi^2=(62-85.9)^2/85.9....=\mathbf{31.4}$
$g.l=(c-1)\cdot(r-1)=(4-1)\cdot(3-1)=6$
$\chi^2_{6;0.05}=12.59$

⇒ 31.4>12.59: si rifiuta l'ipotesi nulla: l'igiene orale è legata al tipo di scuola frequentato

In quale tipo di scuola gli studenti hanno un'igiene orale migliore?

tipo di scuola	**proporzione di studenti con buona igiene orale**		**ES(p)**
sotto la media	62/233=	0.27	0.02895
media	50/119=	0.42	0.04525
sopra la media	80/169=	0.47	0.03841

⇩

A livelli di scuola migliori si accompagna anche una migliore igiene orale

CORREZIONE DI YATES

Quando le **frequenze attese** in una tabella di contingenza sono **inferiori a 5**, la distribuzione teorica di χ^2 non rappresenta bene la distribuzione della statistica χ^2 calcolata sul campione.

Si introduce una correzione nella formula, che diventa:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{t} \frac{\left(|O_{ij} - E_{ij}| - 0.5\right)^2}{E_{ij}}$$

NB frequenze attese **≤ 5** ⟹ raggruppamento in nuove classi

Esercizio 5

Sono state esaminate due aree verdi: dalla più estesa sono state estratte 66 aree, 58 delle quali contenevano alcune particolari specie arboree, mentre nell'area più piccola, tali specie erano presenti in 12 delle 22 unità estratte. Si può affermare che le specie si distribuiscono uniformemente nelle due aree?

Soluzione:

Nella tabella seguente riportiamo le frequenze osservate e le frequenze attese nell'ipotesi di indipendenza.

	numero di unità in cui		
frequenze osservate	**la specie è presente**	**la specie è assente**	
area A	58	8	66
areaB	12	10	22
	70	18	**88**
frequenze attese			
area A	52.5	13.5	66
areaB	17.5	4.5	22
	70	18	**88**

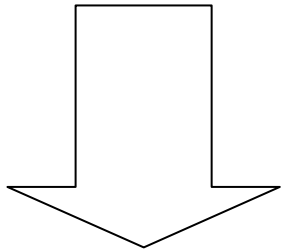
Freq. Attesa <5

Applico la correzione di Yates

$\chi^2=(|58-52.5|-0.5))^2/52.5+(|8-13.5|-0.5)^2/13.5$
$+(|12-17.5|-0.5)^2/17.5+(|10-4.5|-0.5)^2/4.5=$

$= 5^2/(52.5+13.5+17.5+4.5)=25\cdot0.3725=9.312$

$$\chi^2_{1;0.05}=3.84$$



Le specie si distribuiscono con una diversa densità nelle due aree

Memento

- Il χ^2 si utilizza solo per il confronto di frequenze osservate e attese (non per il confronto di tassi, proporzioni ecc

- In genere, quando le freq. attese sono ≤5 si tende a raggruppare in nuove classi per garantire la stabilità della distribuzione

- Se non è possibile riclassificare (es. Tab 2x2) si utilizza la correzione di Yates