

# STATISTICA INFERENZIALE PER VARIABILI QUALITATIVE

La presentazione dei dati per molte ricerche mediche fa comunemente riferimento a frequenze, assolute o percentuali. Osservazioni cliniche conducono sovente a risultati tipo "il 60% degli individui trattati con un farmaco è migliorato rispetto al 47% del gruppo di soggetti di controllo", implicando con ciò un confronto tra i risultati ottenuti per i due gruppi.

Risulta evidente che tali risultati non sono espressi da dati su scala quantitativa e quindi non è possibile fare riferimento alla distribuzione Gaussiana o a quelle del t di Student, ma occorre considerare metodiche specifiche che permettano, anche con tale tipo di dati, di verificare l'ipotesi zero di una differenza casuale tra le frequenze riscontrate.

# La statistica chi-quadrato ( $\chi^2$ )

Variabile statistica semplice (v.s.s.)

**Esempio 1.** C'è parità tra i 2 sessi nei 180 iscritti al corso di laurea in medicina? Si organizza un'indagine su un campione casuale di 80 studenti.  
( $H_0: M=F$ ;  $H_1: M \neq F$ )

I risultati osservati (O) e le attese (A) sono riportati nella tabella.

SESSO	O <sub>1</sub>	A	$\chi^2$ -test	O <sub>2</sub>	$\chi^2$ -test
M	45	40	25/40	50	100/40
F	35	40	25/40	30	100/40
TOT	80	80	50/40	80	200/40
$\chi^2$ g.l.=1			1.25 n.s.		5*
* $p < 0.05$ , risultato del test appena significativo					

## V.S.S. con >2 modalità

Anche una serie empirica può seguire un modello.

**Esempio 2.** 4 campioni di 400 pz. ciascuno vengono sottoposti a  $\neq$  dosaggi di un farmaco. Si riporta il numero osservato di pz guariti ( $O_i$ ) e il numero atteso ( $A_i$ ) per ogni campione ( $C_i$ ).

dove  $H_0$  (modello):  $A_i =$  scala a raddoppio

$\Delta O_i - A_i$  dovuta ad errore?

Dose di farmaco	$O_i$	$A_i$	$\chi^2$ -test
0.5 mg	40	50	100/50
1.0 mg	110	100	100/100
2.0 mg	250	200	2.500/200
4.0 mg	350	400	2.500/400
$\chi^2$	750	750	22.75

# VARIABILI STATISTICHE DOPPIE: CONFRONTO DI 2 CAMPIONI

(Confronto tra due percentuali)

**Esempio 3.** Si abbia un campione di 1020 soggetti diviso in

Fumatori (A):  $n_A=400$       **➔** Prevalenza BCO 30%

Non fumatori (B):  $n_B=620$  **➔** Prevalenza BCO 15%

**Il fumo è “causa” (o fattore di rischio) per la bronchite? ossia il  $\Delta$  (+15%) è statisticamente significativo?**

Tabella di contingenza (2x2)

	BRONCHITE		
FUMO	SI	NO	TOT
SI	120	280	400
NO	93	527	620
TOTALE	213	807	1020

La prevalenza di bronchite risulta statisticamente  $\neq$  tra i fumatori e i non fumatori?

**H0:** La bronchite si sviluppa **indipendentemente** dal fumo;

**H1:** I fumatori sviluppano bronchite più dei non fumatori.

# TASSI DI PREVALENZA x 100 SOGGETTI

$$P_T = \frac{213}{1020} = 20.8\%$$

$$P_F = \frac{120}{400} = 30\%$$

$$P_{NF} = \frac{93}{620} = 15\%$$

- Se ci fosse indipendenza tra fumo e BCO si dovrebbero riscontrare le stesse prevalenze di pazienti con BCO tra i fumatori e i non fumatori.
- Va costruita quindi una tabella le cui frequenze rispondono alla condizione d'indipendenza

# TABELLA TETRACORICA D'INDIPENDENZA

Fattore di rischio	Malattia		TOT
	P (+)	NP (-)	
A (+)	a	b	$n_A(a+b)$
B (-)	c	d	$n_B(c+d)$
TOT	$n_1(a+c)$	$n_2(b+d)$	n

Valori delle frequenze nel caso di indipendenza

$$n_1:n = a:n_A$$



$$a = \frac{n_A n_1}{n}$$

$$n_1:n = c:n_B$$

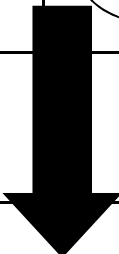


$$c = \frac{n_B n_1}{n}$$

idem per b e d

Tornando all'esempio dell'associazione tra BCO e fumo si ha la tabella delle frequenze attese:

	BRONCHITE CRONICA		
FUMO	SI	NO	TOT
SI	84	316	400
NO	129	491	620
TOTALE	213	807	1020



Es.  $(620 \times 213) / 1020 = 129$ ; per differenza si calcolano le altre tre frequenze interne.

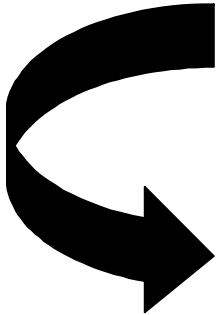
$$\chi^2 = \frac{(120 - 84)^2}{84} + \frac{(280 - 316)^2}{316} +$$
$$+ \frac{(93 - 129)^2}{129} + \frac{(527 - 491)^2}{491} = 32.21$$

**LA FORMULA PER CALCOLARE L'INDICE-TEST  
CHI-QUADRATO**

$$\sum_i \frac{(O_i - A_i)^2}{A_i}$$



Nel caso di tabelle 2x2 si può calcolare il valore del test  $\chi^2$  anche direttamente attraverso la formula seguente:


$$\chi^2 \text{-test} = \frac{(ad - cb)^2 N}{N_1 N_2 N_A N_B}$$

**FORMULA PER IL CALCOLO DEL  $\chi^2$  VALIDA SOLO NEL CASO DI TABELLE TETRACORICHE**

Nel nostro esempio avremo:

$$\chi^2 = \frac{(120 * 527 - 93 * 280)^2 * 1020}{213 * 807 * 400 * 620} = 32.21$$

Valore quasi coincidente a quello calcolato con la precedente formula, quindi

**LE DUE FORMULE DANNO RISULTATI EQUIVALENTI**

Se il campione e 1/10 del precedente si ha:

<b>FUMO</b>	<b>BCO SI</b>	<b>BCO NO</b>	<b>TOTALE</b>
<b>SI</b>	<b>12</b>	<b>28</b>	<b>40</b>
<b>NO</b>	<b>9</b>	<b>53</b>	<b>62</b>
<b>TOTALE</b>	<b>21</b>	<b>81</b>	<b>102</b>

$$\chi^2 = \frac{((12 \cdot 53 - 9 \cdot 28) - 56)^2 \cdot 102}{21 \cdot 81 \cdot 40 \cdot 62} = 2.681$$

L'ipotesi nulla non può essere rifiutata.

## La CORREZIONE di YATES (per la continuità)

La correzione di Yates viene applicata nel caso di tabelle 2x2 che presentino:

- la numerosità complessiva ( $n$ )  $< 200$
- oppure una delle marginali ( $n_A, n_B, n_1, n_2$ )  $< 40$
- comunque  $a, b, c, d > 5$

la correzione si attua con la formula:

$$\chi^2 = \frac{(|ad - cb| - n/2)^2 n}{n_1 n_2 n_A n_B}$$

**Esempio** Si supponga di aver rilevato, su un campione di 36 giovani, la pressione arteriosa e la pratica sportiva.

PRATICA SPORTIVA	IPERTENSIONE ARTERIOSA		
	SI	NO	TOT
SI	7	9	16
NO	14	6	20
TOT	21	15	36

Applichiamo il test del chi-quadrato con la correzione di Yates per la continuità

$$\chi^2 = \frac{((7 \cdot 6 - 14 \cdot 9) - 36 / 2)^2 \cdot 36}{21 \cdot 15 \cdot 20 \cdot 16} = 1.55$$

Il test risulta non significativo dunque l'ipotesi nulla di indipendenza tra la pratica sportiva e l'ipertensione arteriosa viene accettata

FUMO	BCO SI	BCO NO	TOTALE
SI	12	28	40
NO	9	53	62
TOTALE	21	81	102

$$\chi^2 = \frac{((12 \cdot 53 - 9 \cdot 28) - 56)^2 \cdot 102}{21 \cdot 81 \cdot 40 \cdot 62} = 2.681$$

## TEST ESATTO di FISCHER

Viene applicato nel caso in cui in una tabella 2x2 il numero delle osservazioni è minore di 20 o una delle frequenze attese è inferiore a 5. Permette di calcolare direttamente la probabilità esatta.

$$P = \frac{(a+b)! (c+d)! (a+c)! (b+d)!}{a! b! c! d! N!}$$

PRATICA	IPERTENSIONE ARTERIOSA		
	SI	NO	TOT
SPORT			
SI	1	10	11
NO	15	5	20
TOT	16	15	31

$$P_1 = \frac{11! 20! 16! 15!}{1! 10! 15! 5! 31!} = 0.000567$$

PRATICA SPORT	IPERTENSIONE ARTERIOSA		
	SI	NO	TOT
SI	0	11	11
NO	16	4	20
TOT	16	15	31

$$P_0 = \frac{11! 20! 16! 15!}{0! 11! 16! 4! 31!} = 0.000016$$

$$P = 0.000567 + 0.000016 = 0.000583$$

**Altamente significativo.  $P < 0.001$**

Generalizzazione al caso di una tabella di dimensione rxs.

**Esempio 5.**

	Guariti	Migliorati	Non migliorati	Tot.
Farmaco A	21 (15)	15 (17)	7 (11)	43
Farmaco B	12 (18)	24 (22)	18 (14)	54
Tot.	33	39	25	97

33/97=34.02%(GUARITI)

39/97=40.20% (MIGLIORATI)

25/97=25.77% (INSUCCESSI TERAPEUTICI)

21/43=48.8%

15/43=34.9%

12/54=22.2%

18/54=33.3%

Si applica la formula generale per una valutazione complessiva:

$$\chi^2 = \frac{(21-15)^2}{15} + \frac{(12-18)^2}{18} + \frac{(15-17)^2}{17} + \frac{(24-22)^2}{22} + \frac{(7-11)^2}{11} + \frac{(18-14)^2}{14} = 8.23$$



# TEORIA DELLE IPOTESI

$H_0$  = ipotesi zero o ipotesi nulla

le due percentuali (30% e 15%) differiscono per effetto dell'errore di campionamento.

$H_1$  = ipotesi alternativa

le due percentuali non differiscono per effetto dell'errore di campionamento.

il test del  $\chi^2$  consente di saggiare l'ipotesi nulla.

## CONFRONTO TRA PERCENTUALI IN CAMPIONI INDIPENDENTI

Campione 1: n<sub>1</sub>=300    Prevalenza 70%

Campione 2: n<sub>2</sub>=400    Prevalenza 80%

	+	-	TOT
C1	210	90	300
C2	320	80	400
TOT	530	170	700

$$\chi^2 = \frac{(210 \times 80 - 320 \times 90)^2}{530 \times 170 \times 300 \times 400} = 9.32 \quad \mathbf{p < 0.001}$$

Campione 1: n<sub>1</sub>=30    Prevalenza 70%

Campione 2: n<sub>2</sub>=40    Prevalenza 80%

	+	-	TOT
C1	21	9	30
C2	32	8	40
TOT	53	17	70

$$\chi^2 = \frac{(|21 \times 8 - 32 \times 9| - 70/2)^2}{53 \times 17 \times 30 \times 40} = 0.47 \quad \mathbf{n.s.}$$

# TEORIA DELLA VERIFICA DELLE IPOTESI STATISTICHE

Consiste nello stabilire se l'assunzione fatta, si possa considerare esatta o meno, sulla base delle osservazioni condotte su una parte delle unità del collettivo medesimo.

## L'IPOTESI ( $H_0$ ) = ipotesi zero o ipotesi nulla

E' un assunto particolare circa le caratteristiche (i parametri della popolazione. E' una affermazione su eventi "sconosciuti" costruita in modo tale da poter essere verificata mediante un test statistico.

## TEST STATISTICO

E' una tecnica di inferenza statistica, mediante la quale si accetta o rifiuta una certa ipotesi, ad un livello critico di significatività.

## LIVELLO DI SIGNIFICATIVITA'

E' il margine d'errore che siamo disposti a commettere, di solito 5 o 1%, ma più è piccolo e più riduciamo il rischio di rifiutare  $H_0$  quando in realtà è vera.

## FUNZIONE TEST

E' la funzione dei dati campionari di cui si serve un test per portare alla decisione di accettare o respingere  $H_0$ .

## VERIFICA D'IPOTESI

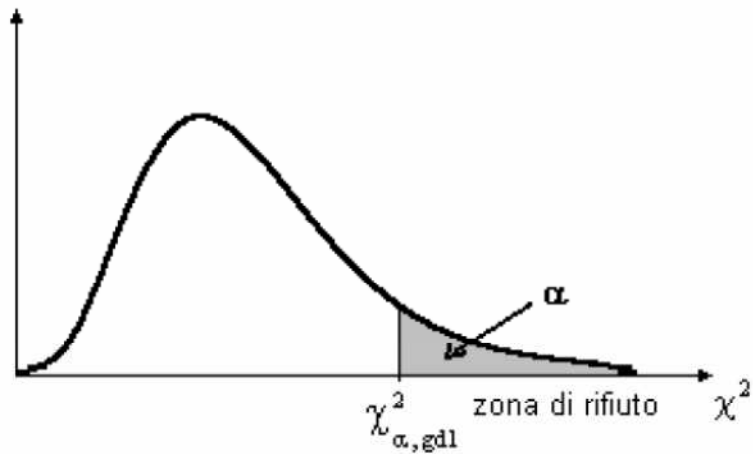
E' una metodologia statistica che basandosi sulle probabilità porta a prendere delle decisioni.

## GRADI DI LIBERTA'

Sono dati, in generale, dal numero delle modalità che la variabile assume meno i vincoli.

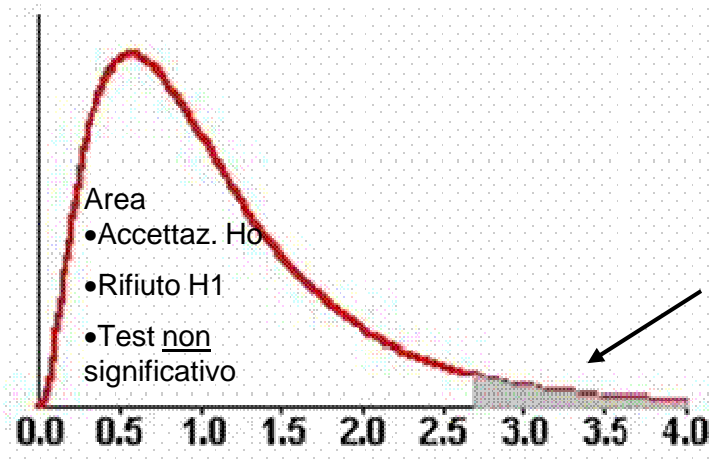
g.l.= $r-1$  per variabili statistiche semplice

g.l.= $(r-1)(c-1)$  per variabili statistiche doppie



<b>d.f.</b>	<b><math>\alpha=0.250</math></b>	<b>0.100</b>	<b>0.050</b>	<b>0.025</b>	<b>0.010</b>	<b>0.005</b>
<b>1</b>	1.3233	2.7055	3.8415	5.0239	6.6349	7.8794
<b>2</b>	2.7726	4.6052	5.9915	7.3778	9.2104	10.5965
<b>3</b>	4.1083	6.2514	7.8147	9.3484	11.3449	12.8381
<b>4</b>	5.3853	7.7794	9.4877	11.1433	13.2767	14.8602
<b>5</b>	6.6257	9.2363	11.0705	12.8325	15.0863	16.7496
<b>6</b>	7.8408	10.6446	12.5916	14.4494	16.8119	18.5475
<b>7</b>	9.0371	12.0170	14.0671	16.0128	18.4753	20.2777

## FUNZIONE TEST



Area

- Accettaz.  $H_0$
- Rifiuto  $H_1$
- Test non significativo

Area

- Rifiuto  $H_0$
- Accettaz.  $H_1$
- Test significativo

**IPOTESI DA**

**VERIFICARE**

D'INDIPENDENZA  
(1)  $H_0: n_{ij}=n'_{ij}$   
 $H_1: n_{ij} \neq n'_{ij}$

DI CONFORMITA'  
O ADATTAMENTO  $H_0: f_o=f_A$   
 $H_1: f_o \neq f_A$

**TEST DI**

**SIGNIFICATIVITA'**

SIGNIFICATIVO → dipendenza tra x e y  
(1)

NON SIGNIFICATIVO → indipendenza

SIGNIFICATIVO → rifiuto il modello  
(s)

NON SIGNIFICATIVO → non rifiuto il modello →  
RISPONDEZZA TRA  
DISTRIBUZIONE CONSTATATA  
E QUELLA TEORICA.