

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 (χ^2)

χ^2 -test (non parametrico)

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

SESSO	O ₁	A	χ^2 -test	O ₂	χ^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
χ^2 g.l.=1			1.25 n.s.	5 *	

=3.84 per $p < 0.05$

VARIABILI STATISTICHE DOPPIE

CONFRONTO DI 2 CAMPIONI

(Confronto tra due percentuali)

Esempio 3. Si abbia un campione di 1020 soggetti diviso in
 Fumatori (A): $n_A=400$ \longrightarrow Prevalenza BCO 30%
 Non fumatori (B): $n_B=620$ \longrightarrow Prevalenza BCO 15%

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

Tabella di contingenza (2x2)

FUMO	BRONCHITE		
	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:

FUMO	BRONCHITE CRONICA		
	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 ALLA TABELLA.

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

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

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

TEORIA SULLE IPOTESI

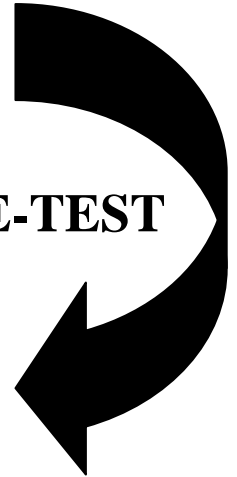
H_0 = IPOTESI ZERO O IPOTESI NULLA

LE DUE PERCENTUALI (30% E 15%) DIFFERISCONO PER EFFETTO DELL'ERRORE DI CAMPIONAMENTO.

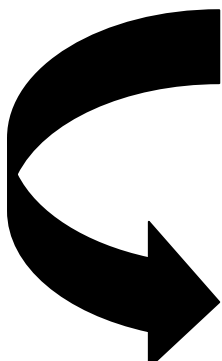
H_1 = IPOTESI ALTERNATIVA

LE DUE PERCENTUALI NON DIFFERISCONO PER EFFETTO DELL'ERRORE DI CAMPIONAMENTO.

IL TEST DEL χ^2 CONSENTE DI SAGGIARE L'IPOTESI NULLA.



Nel caso di tabelle 2x2 si può calcolare il valore del test χ^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 χ^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, ma la seconda è più semplice da calcolare

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

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

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

il test calcolato mediante la formula di Yates è pari a 2.7. quindi l'ipotesi nulla non può essere rifiutata, ma può essere accettata anche se non con certezza, ma con un dato valore probabilistico, in questo caso del 5% (vedere a pagina 13)

MISURE DI RISCHIO

$$\frac{BCO (Fum)}{Fumatori} = \frac{120}{400} 100 = 30 \%$$

$$\frac{BCO (NonFum)}{NonFumatori} = \frac{93}{620} 100 = 15 \%$$

$$RR = \frac{120}{400} : \frac{93}{620} = \frac{120}{400} \frac{620}{93} = \frac{74400}{37200} = 2$$

$$RA = \frac{120}{400} - \frac{93}{620} = (30 - 15)\% = 15 \%$$

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 uno tra $n_A, n_B, n_1, n_2 < 40$

$a, b, c, d > 5$

la correzione si attua con la formula:

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

N.B. Anche per $n > 200$ conviene applicarla

Esempio 4. Si supponga di aver rilevato, su un campione di 36 giovani, la pressione arteriosa e la pratica sportiva.

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

Applichiamo il test χ^2 con la correzione di Yates per la continuità

$$\chi^2 = \frac{(|7 \times 6 - 14 \times 9| - 36/2)^2}{21 \times 15 \times 20 \times 16} = 1.55 \text{ n.s.}$$

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

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

L'ipotesi nulla non può essere rifiutata, ma può essere accettata anche se non con certezza, ma con un dato valore probabilistico.

2.68 è un valore inferiore a quello teorico (tabulato) - riportato alla fine delle diapositive- che per 1 grado di libertà e per una probabilità di errore del 5% è pari a 4.7. tutto ciò ci porta a concludere che il test è risultato non statisticamente significativo, con una probabilità di errore minore del 5% (p-value < 0.05).

Ma nella prima tabella il test era risultato altamente significativo! Dipende anche dalla **dimensione del campione che era 10 volte superiore (N=1.020)**

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 SPORT	IPERTENSIONE ARTERIOSA		
	SI	NO	TOT
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

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

$$P = 0.00567 + 0.000016 = 0.00568$$

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:

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

$$\frac{(18-14)^2}{14} = 8.23$$

CONFRONTO TRA PERCENTUALI IN CAMPIONI INDIPENDENTI

Campione 1: $n_1=300$ Prevalenza 70%

Campione 2: $n_2=400$ 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: $n_1=30$ Prevalenza 70%

Campione 2: $n_2=40$ Prevalenza 80%

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

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

VERIFICA DI IPOTESI

1. 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.

2. L'IPOTESI (H_0)

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 (T.S.)

3. TEST STATISTICO

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

4. 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.

5. FUNZIONE TEST

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

6. VERIFICA D'IPOTESI

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

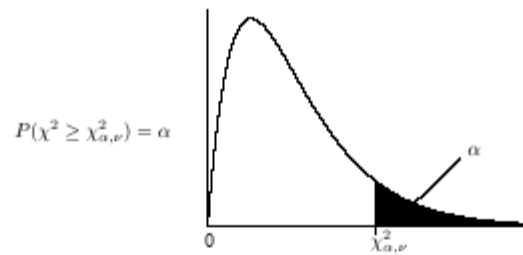
7. GRADI DI LIBERTA'

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

Tavola 3: Valori critici della Distribuzione Chi-Quadrato



ν	0.9999	0.9995	0.999	0.995	α 0.99	0.975	0.95	0.90	0.80
1	1.57E-8	3.93E-7	1.57E-6	3.93E-5	0.0002	0.0010	0.0039	0.0158	0.0642
2	0.0002	0.0010	0.0020	0.0100	0.0201	0.0506	0.1026	0.2107	0.4463
3	0.0052	0.0153	0.0243	0.0717	0.1148	0.2158	0.3518	0.5844	1.0052
4	0.0284	0.0639	0.0908	0.2070	0.2971	0.4844	0.7107	1.0636	1.6488
5	0.0821	0.1581	0.2102	0.4118	0.5543	0.8312	1.1455	1.6103	2.3425
6	0.1723	0.2994	0.3810	0.6757	0.8721	1.2373	1.6354	2.2041	3.0701
7	0.2998	0.4849	0.5985	0.9893	1.2390	1.6899	2.1673	2.8331	3.8223
8	0.4634	0.7104	0.8571	1.3444	1.6465	2.1797	2.7326	3.4895	4.5936
9	0.6611	0.9718	1.1519	1.7349	2.0879	2.7004	3.3251	4.1682	5.3801
10	0.8890	1.2651	1.4787	2.1558	2.5582	3.2470	3.9403	4.8652	6.1791
11	1.1449	1.5870	1.8338	2.6032	3.0535	3.8157	4.5748	5.5778	6.9887
12	1.4281	1.9345	2.2141	3.0738	3.5706	4.4038	5.2260	6.3038	7.8073
13	1.7341	2.3049	2.6172	3.5650	4.1069	5.0087	5.8919	7.0415	8.6339
14	2.0601	2.6966	3.0407	4.0747	4.6604	5.6287	6.5706	7.7895	9.4673
15	2.4084	3.1073	3.4825	4.6009	5.2294	6.2621	7.2609	8.5468	10.3070
16	2.7736	3.5357	3.9417	5.1422	5.8122	6.9077	7.9616	9.3122	11.1521
17	3.1561	3.9800	4.4162	5.6973	6.4077	7.5642	8.6718	10.0852	12.0023
18	3.5559	4.4391	4.9048	6.2648	7.0149	8.2307	9.3904	10.8649	12.8570
19	3.9687	4.9125	5.4067	6.8439	7.6327	8.9065	10.1170	11.6509	13.7158
20	4.3950	5.3978	5.9210	7.4338	8.2604	9.5908	10.8508	12.4426	14.5784
21	4.8342	5.8954	6.4467	8.0336	8.8972	10.2829	11.5913	13.2396	15.4446
22	5.2862	6.4041	6.9829	8.6427	9.5425	10.9823	12.3380	14.0415	16.3140
23	5.7482	6.9240	7.5291	9.2604	10.1957	11.6885	13.0905	14.8480	17.1865
24	6.2231	7.4528	8.0847	9.8862	10.8563	12.4011	13.8484	15.6587	18.0618
25	6.7087	7.9905	8.6494	10.5196	11.5240	13.1197	14.6114	16.4734	18.9397
26	7.1980	8.5374	9.2222	11.1602	12.1982	13.8439	15.3792	17.2919	19.8202
27	7.6997	9.0929	9.8029	11.8077	12.8785	14.5734	16.1514	18.1139	20.7030
28	8.2115	9.6558	10.3907	12.4613	13.5647	15.3079	16.9279	18.9392	21.5880
29	8.7303	10.2266	10.9861	13.1211	14.2564	16.0471	17.7084	19.7677	22.4751
30	9.2559	10.8040	11.5876	13.7867	14.9535	16.7908	18.4927	20.5992	23.3641
35	11.9929	13.7879	14.6881	17.1917	18.5089	20.5694	22.4650	24.7966	27.8359
40	14.8820	16.9058	17.9166	20.7066	22.1642	24.4331	26.5093	29.0505	32.3449
45	17.8922	20.1361	21.2509	24.3110	25.9012	28.3662	30.6123	33.3504	36.8844
50	21.0077	23.4611	24.6736	27.9908	29.7067	32.3574	34.7642	37.6886	41.4492
55	24.2133	26.8650	28.1731	31.7349	33.5705	36.3981	38.9581	42.0596	46.0356
60	27.5006	30.3393	31.7381	35.5344	37.4848	40.4817	43.1880	46.4589	50.6406
70	34.2581	37.4671	39.0358	43.2753	45.4417	48.7575	51.7393	55.3289	59.8978
80	41.2407	44.7917	46.5197	51.1719	53.5400	57.1532	60.3915	64.2778	69.2070
90	48.4095	52.2768	54.1559	59.1963	61.7540	65.6466	69.1260	73.2911	78.5584
100	55.7202	59.8946	61.9182	67.3275	70.0650	74.2219	77.9294	82.3581	87.9453
150	93.9492	99.4617	102.1127	109.1423	112.6676	117.9846	122.6918	128.2750	135.2625
200	134.0154	140.6591	143.8420	152.2408	156.4321	162.7280	168.2785	174.8353	183.0028

Tabola 3 (segue): Valori critici della Distribuzione Chi-Quadrato

ν	α								
	0.20	0.10	0.05	0.025	0.01	0.005	0.001	0.0005	0.0001
1	1.6424	2.7055	3.8415	5.0239	6.6349	7.8794	10.8274	12.1153	15.1343
2	3.2189	4.6052	5.9915	7.3778	9.2104	10.5965	13.8150	15.2014	18.4247
3	4.6416	6.2514	7.8147	9.3484	11.3449	12.8381	16.2660	17.7311	21.1040
4	5.9886	7.7794	9.4877	11.1433	13.2767	14.8602	18.4662	19.9977	23.5064
5	7.2893	9.2363	11.0705	12.8325	15.0863	16.7496	20.5147	22.1057	25.7507
6	8.5581	10.6446	12.5916	14.4494	16.8119	18.5475	22.4575	24.1016	27.8527
7	9.8032	12.0170	14.0671	16.0128	18.4753	20.2777	24.3213	26.0179	29.8814
8	11.0301	13.3616	15.5073	17.5345	20.0902	21.9549	26.1239	27.8674	31.8268
9	12.2421	14.6837	16.9190	19.0228	21.6660	23.5893	27.8767	29.6669	33.7247
10	13.4420	15.9872	18.3070	20.4832	23.2093	25.1881	29.5879	31.4195	35.5572
11	14.6314	17.2750	19.6752	21.9200	24.7250	26.7569	31.2635	33.1382	37.3647
12	15.8120	18.5493	21.0261	23.3367	26.2170	28.2997	32.9092	34.8211	39.1306
13	16.9848	19.8119	22.3620	24.7356	27.6882	29.8193	34.5274	36.4768	40.8735
14	18.1508	21.0641	23.6848	26.1189	29.1412	31.3194	36.1239	38.1085	42.5752
15	19.3107	22.3071	24.9958	27.4884	30.5780	32.8015	37.6978	39.7173	44.2596
16	20.4651	23.5418	26.2962	28.8453	31.9999	34.2671	39.2518	41.3077	45.9255
17	21.6146	24.7690	27.5871	30.1910	33.4087	35.7184	40.7911	42.8808	47.5591
18	22.7595	25.9894	28.8693	31.5264	34.8052	37.1564	42.3119	44.4337	49.1853
19	23.9004	27.2036	30.1435	32.8523	36.1908	38.5821	43.8194	45.9738	50.7873
20	25.0375	28.4120	31.4104	34.1696	37.5663	39.9969	45.3142	47.4977	52.3832
21	26.1711	29.6151	32.6706	35.4789	38.9322	41.4009	46.7963	49.0096	53.9599
22	27.3015	30.8133	33.9245	36.7807	40.2894	42.7957	48.2676	50.5105	55.5244
23	28.4288	32.0069	35.1725	38.0756	41.6383	44.1814	49.7276	51.9995	57.0668
24	29.5533	33.1962	36.4150	39.3641	42.9798	45.5584	51.1790	53.4776	58.6071
25	30.6752	34.3816	37.6525	40.6465	44.3140	46.9280	52.6187	54.9475	60.1360
26	31.7946	35.5632	38.8851	41.9231	45.6416	48.2898	54.0511	56.4068	61.6666
27	32.9117	36.7412	40.1133	43.1945	46.9628	49.6430	55.4751	57.8556	63.1660
28	34.0266	37.9159	41.3372	44.4608	48.2782	50.9936	56.8918	59.2990	64.6561
29	35.1394	39.0875	42.5569	45.7223	49.5878	52.3355	58.3006	60.7342	66.1524
30	36.2502	40.2560	43.7730	46.9792	50.8922	53.6719	59.7022	62.1600	67.6230
35	41.7780	46.0588	49.8018	53.2033	57.3420	60.2746	66.6192	69.1975	74.9253
40	47.2685	51.8050	55.7585	59.3417	63.6908	66.7660	73.4029	76.0963	82.0551
45	52.7288	57.5053	61.6562	65.4101	69.9569	73.1660	80.0776	82.8734	89.0704
50	58.1638	63.1671	67.5048	71.4202	76.1538	79.4898	86.6603	89.5597	95.9713
55	63.5772	68.7962	73.3115	77.3804	82.2920	85.7491	93.1671	96.1607	102.7735
60	68.9721	74.3970	79.0820	83.2977	88.3794	91.9518	99.6078	102.6971	109.4967
70	79.7147	85.5270	90.5313	95.0231	100.4251	104.2148	112.3167	115.5766	122.7443
80	90.4053	96.5782	101.8795	106.6285	112.3288	116.3209	124.8389	128.2636	135.7728
90	101.0537	107.5650	113.1452	118.1359	124.1162	128.2987	137.2082	140.7804	148.6198
100	111.6667	118.4980	124.3421	129.5613	135.8069	140.1697	149.4488	153.1638	161.3297
150	164.3492	172.5812	179.5806	185.8004	193.2075	198.3599	209.2652	213.6135	223.1209
200	216.6088	226.0210	233.9942	241.0578	249.4452	255.2638	267.5388	272.4220	283.0448